APPENDIX 2

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:
Eva KONTSEKOVA
Peter FILIPCIK

Serial No.: 10/521,049

Filed: November 1, 2005

For: TRANSGENIC ANIMAL EXPRESSING ALZHEIMER'S TAU PROTEIN

Group Art Unit: 1633

Examiner: Leavitt, Maria Gomez

Atty. Dkt. No.: SONN:066US

CERTIFICATE OF ELECTRONIC SUBMISSION

DATE OF SUBMISSION: Jan. 10, 2007

FILIPCIK DECLARATION UNDER 37 C.F.R. § 1.132

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

I, Peter Filipcik, declare that:

1. I am a co-inventor of the above-referenced patent application. I am also an employee of Axon Neuroscience, the assignee of the above-referenced application. A copy of my *Curriculum Vitae* is attached as Exhibit 1.

2. I am a co-author of the publication Zilka *et al.*, "Truncated tau from sporadic Alzheimer's disease suffices to drive neurofibrillary degeneration in vivo," *FEBS Letters* 580:3582-3588 (2006) (hereinafter, the "Zilka reference").

3. It is my understanding that the Examiner in charge of the above-captioned application has advanced an enablement rejection against claims 17-33. I am supplying this declaration to provide additional evidence of the enablement of the present claims. In particular, this

declaration provides additional data on transgenic rat line #318, which is the same transgenic rat line #318 described in the present patent application, demonstrating that a transgenic animal having a DNA construct coding for N- and C-terminally truncated tau molecules according to the present invention exhibits phenotypes that make it a suitable model for Alzheimer's disease.

4. Attached as Exhibit 2 is Zilka *et al.*, "Truncated tau from sporadic Alzheimer's disease suffices to drive neurofibrillary degeneration in vivo," *FEBS Letters* 580:3582-3588 (2006). The Zilka reference describes the generation of and studies on transgenic rat line #318. Transgenic rat line #318 is the same transgenic rat line #318 described in the specification of the present patent application. *See, e.g.*, Specification, p. 22, first full paragraph; p. 23, first full paragraph; and Fig. 3C.

5. According to the teachings in the present specification, the DNA constructs used for transgenic animal preparation in the Zilka reference are characterized by the following features: (1) the cDNA molecules are truncated at least 30 nucleotides downstream of the start codon and truncated at least the 30 nucleotides upstream of the stop codon of the full-length tau cDNA sequence coding for 4-repeat and 3-repeat tau protein; (2) the cDNA molecule comprises SEQ ID No. 9; (3) and the DNA construct encodes a protein, which has neurofibrillary (NF) pathology producing activity when expressed in brain cells.

6. The transgene construct used in the generation of transgenic rat lines #318 and #72 was prepared by ligation of a cDNA coding for human tau protein truncated at amino acid positions 151-391 into the mouse Thy-1 gene downstream of the brain promoter/enhancer sequence. *Zilka*, p. 3582, col. 2. Transgenic rat line #318 is the same rat line described

in the present patent application (*see e.g.*, Example 2). It should be noted that the numbering of the amino acids of the tau protein in the Zilka reference is based on tau isoform 40, whereas the numbering in the present patent application is based on tau isoform 43. Tau isoform 40 contains an extra insert of 58 amino acids (174 nucleotides) in the N-terminus of the protein. Thus, the truncated tau protein numbered amino acids 151-391 in the Zilka reference is the same as a truncated tau protein numbered amino acids 93-333 based on the numbering in the patent application. Using the numbering in the patent application, amino acids 93-333 correspond to nucleotides 279-999. Thus, the truncated tau cDNA molecule used to generate rat line #318 is truncated at least 30 nucleotides downstream of the start codon and truncated at least the 30 nucleotides upstream of the stop codon of the full-length tau cDNA sequence coding for 4-repeat and 3-repeat tau protein; and the truncated tau cDNA molecule comprises SEQ ID NO: 9 (nucleotides 741-930).

7. The transgenic DNA was linearized by cleavage with EcoRI, and the vector sequences were removed prior to microinjection. *Zilka*, p. 3582, col. 2. Transgenic rats were generated by pronuclear injection of one-day old SHR rat embryos. *Id.* Founders were screened by PCR using Thy-1-specific and human tau-specific primers. *Id.* Two independent transgenic founder lines, #318 and #72, that stably expressed human truncated tau were obtained. *Zilka*, p. 3582, col. 2 to p. 3583, col. 1.

8. The Zilka reference also describes the generation of transgenic rat line #72, which was created using the same transgene construct and the same SHR background as used in the generation of transgenic rat line #318. *See e.g.*, Zilka, paragraph spanning pp 3582-3583.

9. As described in the present specification transgenic rat line #318 exhibits neurofibrillary (NF) pathology producing activity when expressed. For example, Fig. 6 shows the detection of intracellular inclusions and neurofibrillary filaments using silver staining in the neurons of the central nervous system of transgenic rats, whereas wild-type rats did not show these structures in the homologous brain area. Figs. 7 and 8 show the detection of neurofibrillary tangles in the central nervous system of transgenic rats using the pan-tau monoclonal antibody DC 25 and the monoclonal antibody PHF-1, respectively. Additionally, Fig. 10 shows a comparison of neurofibrillary tangles detected by Gallyas silver technique (Fig. 10A and Fig. 10C) and also by immunohistochemistry (Fig. 10E) in AD diseased human brain, to the equivalent pathological structures observed in the transgenic rat of present invention. The observation of neurofibrillary pathology in transgenic rat line #318 described in the present specification was confirmed by the studies described in the Zilka reference in which transgenic rat lines #318 was shown to exhibit neurofibrillary pathology. *Zilka*, p. 3582-3583 and Fig. 3. Thus, the studies presented in the present patent application and in the Zilka reference demonstrate that the DNA construct used to make transgenic rat line #318 encodes a protein, which has neurofibrillary (NF) pathology producing activity when expressed in brain cells.

10. The evidence discussed above demonstrates that transgenic rat line #318 contains a DNA construct having a cDNA molecule coding for N- and C-terminally truncated tau molecules having the following features: (1) the cDNA molecule is truncated at least 30 nucleotides downstream of the start codon and truncated at least the 30 nucleotides upstream of the stop codon of the full-length tau cDNA sequence coding for 4-repeat and 3-repeat tau protein; (2) the cDNA molecule comprises SEQ ID No. 9; and (3) the DNA

construct encodes a protein, which has neurofibrillary (NF) pathology producing activity when expressed in brain cells.

11. In additional studies with transgenic rat line #318 performed at Axon Neuroscience, phenotypes including cognitive impairment, oxidative stress, metabolic (energy) stress, and phosphorylation have been observed. For example, a statistically significant cognitive deficit was measured in transgenic rat line #318 as compared to non-transgenic litter mates in a water maze test. Exhibit 3, Fig. 1. As shown in Fig. 2 of Exhibit 3, transgenic rat line #318 showed increased oxidative stress as a consequence of the pathological cascade initiated by transgene expression. As shown in Fig. 3 of Exhibit 3, the kinetic measurement of the creatine kinase reaction showed that the constant rate values of the brain specific creatine kinase was significantly reduced in transgenic rat line #318 indicating energy stress. In addition, western blot analysis showed strong AD-like phosphorylation pattern of tau protein in transgenic rat line #318. Exhibit 3, Fig. 5.

12. The observed phenotypes described above demonstrate that this transgenic animal is a suitable model for Alzheimer's disease.

13. Furthermore, in addition to the transgenic rat lines #318 and #72, which were generated in the SHR genetic background, the same DNA construct was introduced into the Wistar rat genetic background. The transgenic rat line in the Wistar background exhibited the same neurofibrillary pathology phenotype as the transgenic rat lines in the SHR background. This result indicates that the observed phenotype is associated with the expression of the truncated tau protein and not with the genetics of any particular rat line.

14. As noted in the specification (p. 12, first full paragraph) and in the Zilka reference (p. 3582, col. 2) the Thy-1 promoter was used for the expression of truncated tau. This

Thy-1 promoter is derived from mice and therefore, the constructs would be expected to work for a mouse model in addition to the rat models already tested. Furthermore, sequencing of the rat genome has revealed a high homology between genomes of the rat and the mouse. *See* Exhibit 4 (Rat Genome Sequencing Project Consortium, *Nature* 428:493-521 (2004), particularly Fig. 7), and tau protein exhibits high phylogenetic conservation across a variety of species. There are examples in neurobiology showing that the identical or very homologous gene constructs are responsible for very similar phenotypes in transgenic animals of different species. For example, expression of mutated SOD in rats and mice have produced a very similar phenotype (Gurney *et al.*, *Science*, 264:1772-1774 (1994) (Exhibit 5); Howland *et al.*, *PNAS*, 99(3):1604-1609 (2002) (Exhibit 6)). As a further example, similar results were obtained in modeling Huntington disease in mice and rats (von Horsten *et al.*, *Human Molecular Genetics*, 12(6):617-624 (2003) (Exhibit 7); Bates *et al.*, *Human Molecular Genetics*, 6(10):1633-1637 (1997) (Exhibit 8); Mangiarini *et al.*, *Cell*, 87(3):493-506 (1996) (Exhibit 9)). In view of these observations, it can be expected that the same phenotype as observed in the transgenic rat can also be observed in transgenic mice expressing the same gene construct.

15. I hereby declare that all statements made herein of my knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

8. 1. 2007
Date

Peter Filipcik

EXHIBIT 1

CURICULUM VITAE

NAME: RNDr. Peter Filipcik, PhD

BORN: June 26, 1962

CITIZENSHIP: Slovakia

ADDRESS: Podhaj 3, Lamac, 84103 Bratislava,
e-mail: peter.filipcik@savba.sk

EDUCATION:

June 1995 PhD	Slovak Academy of Sciences, Bratislava, Slovakia
June 1986 RNDr.	Comenius University, Faculty of Natural Sciences in Bratislava, Slovakia

EMPLOYMENT:

1996 – pres	Senior scientist - Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia (part time)
2001 – pres	Senior scientist - Axon Neuroscience GmbH, Vienna, Austria
1986 - 1996	Research assistant, Institute of Experimental Endocrinology, Slovak Academy of Sciences, Bratislava, Slovakia
2000 - 2001	University of Vienna, Vienna, Austria
1998 - 2000	Visiting scientist at the CCRI, St. Anna Children Hospital, Vienna, Austria
1995 - 1996	Research associate, Dept. of Pharmacol., University of Minnesota, Minneapolis, USA
1993 - 1994	Research assistant, Dept. of Chem. Pharmacol., University of Tokyo, Japan

INTERNATIONAL COURSES AND MEETINGS ATTENDED (selection):

1990	"3rd European Congress on Cell Biology", Florence, Italy
1993	"The Radioisotopes in Biological Research", The Univ. of Tokyo, Tokyo, Japan
1993	"5th Inter-Department Meeting on Chemical Pharmacol.", Seoul, South Korea
1998	"6th Int. Conf. on Alzheimer's Disease and Related Disorders, Amsterdam, Netherlands
2001	"Ageing and Dementia - Current and future concepts", Graz, Austria
2003	In Vitro Human Cell Systems Enabling Drug Discovery , London, UK
2004	"9th International Conference on Alzheimers Disease and Related Disorders", Philadelphia, Pennsylvania
2005	Molecular Medicine Triconference, CHI, San Francisco, California, USA
2006	"10th International Conference on Alzheimers Disease", Madrid, Spain

MEMBERSHIP OF LEARNED SOCIETIES:

1997	Slovak Immunological Society
1996	The Slovak Alzheimer Society
2005	The Slovak Neuroscience Society

PUBLICATION ACTIVITY:

Author and co-author of 21 scientific papers, 2 patents

Bratislava 6. 12. 2006

List of publications:

Filipcik P, Cente M, Ferencik M, Hulin I, Novak M. The role of oxidative stress in the pathogenesis of Alzheimer's disease. Bratisl Lek Listy. 2006, 107 (9-10), 384-394

Pevalova M, **Filipcik P**, Novak M, Avila J, Iqbal K. Post-translational modifications of tau protein Bratisl Lek Listy 2006; 107 (9-10), 346-353

*Cente M, ***Filipcik P**, Pevalova M, Novak M. Expression of a truncated tau protein induces oxidative stress in a rodent model of tauopathy. Eur J Neurosci. 2006 Aug;24(4):1085-90.

*Zilka N, ***Filipcik P**, Koson P, Fialova L, Skrabana R, Zilkova M, Rolkova G, Kontsekova E, Novak M. Truncated tau from sporadic Alzheimer's disease suffices to drive neurofibrillary degeneration in vivo. FEBS Lett. 2006 Jun 26;580(15):3582-8.

Soltys K, Rolkova G, Vechterova L, **Filipcik P**, Zilka N, Kontsekova E, Novak M. First insert of tau protein is present in all stages of tau pathology in Alzheimer's disease. Neuroreport. 2005 Oct 17;16(15):1677-81.

Matuskova M, Csokova N, **Filipcik P**, Hanusovska E, Bires J, Cabadaj R, Kontsek P, Novak M. First confirmed sheep scrapie with A136R154Q171 genotype in Slovakia. Acta Virol. 2003;47(3):195-8.

Lion T, Daxberger H, Dubovsky J, **Filipcik P**, Fritsch G, Printz D, Peters C, Matthes-Martin S, Lawitschka A, Gadner H. Analysis of chimerism within specific leukocyte subsets for detection of residual or recurrent leukemia in pediatric patients after allogeneic stem cell transplantation. Leukemia. 2001 Feb;15(2):307-10. No abstract available.

Cattaneo A, Capsoni S, Margotti E, Righi M, Kontsekova E, Pavlik P, **Filipcik P**, Novak M. Functional blockade of tyrosine kinase A in the rat basal forebrain by a novel antagonistic anti-receptor monoclonal antibody. J Neurosci. 1999 Nov 15;19(22):9687-97.

Brtko J, **Filipcik P**, Hudecova S, Brtkova A, Bransova J. Nuclear all-trans retinoic acid receptors: in vitro effects of selenium. Biol Trace Elem Res. 1998 Apr-May;62(1-2):43-50.

Filipcik P, Strbak V, Brtko J. Thyroid hormone receptor occupancy and biological effects of 3,5,3,-L-triiodothyronine (T3) in GH4C1 rat pituitary tumour cells. Physiol Res. 1998;47(1):41-6.

Wei LN, Lee CH, **Filipcik P**, Chang L. Regulation of the mouse cellular retinoic acid-binding protein-I gene by thyroid hormone and retinoids in transgenic mouse embryos and P19 cells. J Endocrinol. 1997 Oct;155(1):35-46.

Nikodemova M, Weismann P, **Filipcik P**, Mraz P, Greer MA, Strbak V. Both iso- and hyperosmotic ethanol stimulate release of hypothalamic thyrotropin-releasing hormone despite opposite effect on neuron volume. Neuroscience. 1997 Oct;80(4):1263-9.

Filipcik P, Brtko J. [The basis for the variable effects of thyroid hormones] Cesk Fysiol. 1996 Mar;45(1):13-20. Slovak.

Brtko J, **Filipcik P**, Hudecova S, Strbak V, Brtkova A. In vitro effects of sodium selenite on nuclear 3,5,3'-triiodothyronine (T3) receptor gene expression in rat pituitary GH4C1 cells. Biol Trace Elem Res. 1995 May;48(2):173-83.

Filipcik P, Saito H, Katsuki H. 3,5,3'-L-triiodothyronine promotes survival and axon elongation of embryonic rat septal neurons. Brain Res. 1994 May 30;647(1):148-52.

Brtko J, **Filipcik P**. Effect of selenite and selenate on rat liver nuclear 3,5,3'-triiodothyronine (T3) receptor. Biol Trace Elem Res. 1994 Apr-May;41(1-2):191-9.

Brtko J, Knopp J, **Filipcik P**, Baker ME. Effect of protease inhibitors and substrates on 3,5,3'-triiodothyronine binding to rat liver nuclear receptors. Endocr Regul. 1992 Sep;26(3):127-31.

Knopp J, Brtko J, Filipcik P. Effect of triiodothyronine on rat liver polysome profiles and translational activity of mRNA after partial hepatectomy. Endocr Regul. 1992 Jun;26(2):67-72.

Brtko J, **Filipcik P**, Knopp J, Sedlakova V, Rauova L. Thyroid hormone responsiveness of the L1210 murine leukemia cell line. Acta Endocrinol (Copenh). 1992 Apr;126(4):374-7.

Filipcik P, Brtko J, Rauova L, Sedlakova V. Distribution of triiodothyronine nuclear receptors during the cell cycle in mouse leukemia cells. Folia Biol (Praha). 1992;38(6):332-9.

Filipcik P, Brtko J, Knopp J. [Cell lines in experimental endocrinology] Bratisl Lek Listy. 1990 Apr;91(4):278-83. Slovak.

---

*Equal contribution.

NEUROBIOLOGY OF AGING supplement:

Filipcik, P; Pevalova, M; Smrzka, O; Novak, M. Neuronal assay based on developmentally inducible expression of Alzheimer's tau, designed for screening of AD therapeutics. NEUROBIOLOGY OF AGING, JUL 2004, 25. Suppl. 2, S265

Pevalova, M; **Filipcik, P**; Mederlyova, A; Cente, M; Smrzka, O; Novak, M Hyperphosphorylation and oxidative stress as early changes in axon's new AD rat model. NEUROBIOLOGY OF AGING, JUL 2004, 25, Suppl. 2, S264

Cente, M; **Filipcik, P**; Hanusovska, E; Zilka, N; Novak, M Onset and intensity of AD changes in transgenic rat expressing Alzheimer specific Tau protein correlates with gene dosage. NEUROBIOLOGY OF AGING, JUL 2004, 25 Suppl. 2, S239

Hrnkova, M; Zilka, N; **Filipcik, P**; Novak, M Cognitive deficit and progressive motor impairment in AD rat model, NEUROBIOLOGY OF AGING, JUL 2004, 25, Suppl. 2, S233

Koson, P; Zilka, N; **Filipcik, P**; Novak, M Neuronal loss in selected brain areas of a new transgenic AD rat model estimated with unbiased stereological methods, NEUROBIOLOGY OF AGING, JUL 2004, 25 Suppl. 2, S249, S250.

Zilka, N; Csokova, N; Vechterova, L; Skrabanova, M; Hrnkova, **M; Filipcik, P**; Novak, M. Staging of neuropathological changes in axon's novel transgenic AD rat model is linked to a lethal phenotype. NEUROBIOLOGY OF AGING, JUL 2004, 25. Suppl. 2, S255

EXHIBIT 2

FEBS Letters 580 (2006) 3582 3588

Truncated tau from sporadic Alzheimer's disease suffices to drive neurofibrillary degeneration in vivo

Norbert Zilka[a,1], Peter Filipcik[a,1], Peter Koson[a], Lubica Fialova[a], Rostislav Skrabana[a], Monika Zilkova[a], Gabriela Rolkova[a], Eva Kontsekova[a], Michal Novak[a,b,*]

[a] *Axon Neuroscience GmbH, Rennweg 95b, 1030 Vienna, Austria*
[b] *Institute of Neuroimmunology, Slovak Academy of Sciences, Dubravska 9, 845 10 Bratislava, Slovak Republic*

Received 20 April 2006; revised 5 May 2006; accepted 8 May 2006

Available online 22 May 2006

Edited by Jesus Avila

Abstract **Truncated tau protein is the characteristic feature of human sporadic Alzheimer's disease. We have identified truncated tau proteins conformationally different from normal healthy tau. Subpopulations of these structurally different tau species promoted abnormal microtubule assembly in vitro suggesting toxic gain of function. To validate pathological activity in vivo we expressed active form of human truncated tau protein as transgene, in the rat brain. Its neuronal expression led to the development of the neurofibrillary degeneration of Alzheimer's type. Furthermore, biochemical analysis of neurofibrillary changes revealed that massive sarcosyl insoluble tau complexes consisted of human Alzheimer's tau and endogenous rat tau in ratio 1:1 including characteristic Alzheimer's disease (AD)-specific proteins (A68). This work represents first insight into the possible causative role of truncated tau in AD neurofibrillary degeneration in vivo.**

Keywords: Alzheimer's disease; Truncated tau; Microtubule assembly; Neurofibrillary degeneration; Sarcosyl insoluble tau; Tau cascade

1. Introduction

Neurofibrillary structures in Alzheimer's disease are principally composed of hyperphosphorylated tau [1–4] and truncated forms of tau protein [2,5,6]. It has been shown that truncation is closely associated with Alzheimer's disease (AD)-typical conformational changes of the tau protein [5–10]. We have hypothesized that truncation could play major role in AD tau pathology [11]. This hypothesis originated from finding that AD-specific monoclonal antibody 423 (mAb 423) recognizes truncated tau species in the core of paired helical filaments (PHF) of Alzheimer's disease [5,6,12,13]. Furthermore, mAb DC11, raised against sporadic AD-brain derived tau extracts, recognized all and only those tau proteins that were truncated at the N-terminus or at both, the N- and C- termini [8]. Truncated tau proteins from sporadic cases of human AD, recognized by mAb DC11 ("Tau DC11 state"), were further tested in vitro for their potency to promote microtubule assembly. The subpopulations of these truncated tau species induced abnormal microtubule assemblies, suggesting toxic gain of function. In order to elucidate the role of truncated tau in AD tau cascade we used truncated tau that was the most active in promotion of abnormal microtubule assembly, as a transgene in the rat brain. Rats displayed massive neurofibrillary structures induced by expressed human truncated tau. This is for the first time shown that truncated human tau could serve as a driving force in neurodegeneration of AD type in vivo.

*Corresponding author. Fax: +421 2 54774276.
E-mail address: Michal.Novak@savba.sk (M. Novak).

[1] These authors contributed equally.

Abbreviations: AD, Alzheimer's disease; mAb, monoclonal antibody; NFT, neurofibrillary tangle; NT, neuropil threads; OD, optical density; PHF, paired helical filaments; SDS–PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis

2. Materials and methods

2.1. Preparation, expression and purification of tau proteins

The preparation of cDNA coding for human tau isoforms and truncated tau proteins was described elsewhere [6]. All DNA constructs were cloned in pET17 vector (Novagen) through *NdeI–EcoRI* restriction sites. Integrity of each construct was verified by DNA sequence analysis (ABI Prism 377DNA Sequencer, Perkin–Elmer). Tau proteins were expressed in *Escherichia coli* and purified from bacterial lysates by ion-exchange chromatography [14]. The protein concentration was determined by BCA kit (Pierce, USA).

2.2. Microtubule assembly

Tubulin for microtubule assembly assay was isolated from pig brains, using reversible assembly purification method [15]. Assay mixtures contained 1 mg/ml tubulin, 1 mM GTP, recombinant tau proteins (0.2 mg/ml) in assembly buffer (100 mM Pipes, pH 6.9; 1 mM $MgSO_4$ and 2 mM EGTA). After gentle and rapid mixing, the samples were pipetted into quartz microcuvettes and equilibrated at 37 °C in a thermostatically controlled spectrophotometer (Beckman Coulter). The turbidity was continuously monitored at 340 nm for a period of 5 min. For electron microscopy samples were fixed with 1% glutaraldehyde, put on the formvar/carbon coated 400 mesh copper grid (Agar Scientific, UK) and stained with 1% aqueous uranyl acetate.

2.3. Preparation of transgene construct and generation of transgenic rats

The transgene construct was prepared by ligation of a cDNA coding for human tau protein truncated at amino acid positions 151–391, into the mouse Thy-1 gene downstream of the brain promoter/enhancer sequence. The original Thy-1 gene sequence coding for exons II–IV, together with thymus enhancer sequence was replaced by the cDNA. Transgenic DNA was linearized by cleavage with EcoRI. Vector sequences were removed prior to microinjection. Transgenic rats were generated by pronuclear injection of one-day old SHR rat embryos. Founders were double screened by PCR using Thy-1-specific and human tau-specific primers amplifying START and STOP codon flanking sequences. The rat endogenous tau sequence was used as an internal amplification control. Two independent transgenic founder

0014-5793/$32.00
doi:10.1016/j.febslet.2006.05.029

lines (#318 and #72) that stably expressed human truncated tau were engineered and displayed similar phenotype. The studies described below were performed using the line #318. The expression level of total human truncated tau in the transgenic rats was determined by Western blot analysis using protein extracts from different areas of brain and spinal cord.

2.4. Monoclonal antibodies

HT7 (Innogenetics, Belgium) recognizes residues 159–163 of human tau, AT8 (Innogenetics) recognizes phosphoserine 202 and 205, AT180 (Innogenetics) recognizes phosphothreonine 231, PHF 1 (a kind gift from Dr. Peter Davies) recognizes phosphoserine 396 and 404. As a control we have used pan tau antibody DC25 recognizing residues 347–354 (Axon Neuroscience, Austria).

2.5. Histology and immunohistochemistry

Animals were perfused transcardially with 4% paraformaldehyde in 0.1 M phosphate buffered saline, pH 7.2, and the tissues were post-fixed after perfusion and then cut on cryotome or embedded in paraffin and cut on microtome. Immunohistochemistry and histopathology were performed on 50 μm free-floating and 8 μm paraffin embedded sections. Tissue sections were immunostained using the standard avidin–biotin–peroxidase method. The modified Gallyas silver iodide, Congo red and Thioflavin S staining methods were utilized to demonstrate mature neurofibrillary pathology in neurons [16,17]. Sections were examined with an Olympus BX51 and Zeiss Axiovert 200 microscopes.

2.6. Stereological analysis

The quantified parameters were neuronal and neurofibrillary tangle (NFT) density. The left brain stems of 10 months old transgenic males were sectioned on cryostat in the frontal plane. The rostral part of the gigantocellular reticular nucleus was selected as a representative region of the reticular formation of the brain stem. NFTs were immunohistochemically visualized using mAb AT8 and thereafter the sections were counterstained with cresyl violet. The optical disector principle was applied [18], particles (neurons, NFTs) were counted and numerical densities per mm^3 were calculated. The obtained results were corrected to the number weighted final section thickness [19] to eliminate any possible bias in the data due to shrinkage of the sections during histological processing. The study was realized with the aid of a computer-based stereological system (StereoInvestigator, MicroBrightField, USA).

2.7. Extraction of sarcosyl insoluble tau

Sarcosyl insoluble tau was isolated from brain tissues of 3–12 months old rats based on the modified method of Greenberg and Davies [20]. Approximately 2 g of brain tissue was homogenized in 10 volumes of buffer (10 mM Tris, 0.8 M NaCl, 1 mM EGTA and 10% sucrose, pH 7.4) and centrifuged at $27200 \times g$ for 20 min. The supernatant was adjusted to 1% (w/v) *N*-lauroylsarcosine and incubated 1 h at RT. After the incubation supernatant was spun at $123000 \times g$ for 1 h at RT. Resulted pellet was resuspended in small volume of phosphate-buffered saline and analysed in Western blot and throughout the paper is designated as P2.

2.8. Western blotting

Sarcosyl insoluble tau proteins purified from brains were analyzed on 5–20% SDS–PAGE gradient gel and Western blot as described previously [14]. Enhanced chemiluminescence developed Western blot was digitalized with LAS3000 CCD imaging system (Fujifilm, Japan). Densitometric data analysis and relative quantification of Western blot record were performed by AIDA Biopackage (Raytest, Germany) as described [14].

3. Results

3.1. Truncated tau protein (t151-391) induces abnormal assembly of microtubules in vitro

Monoclonal antibody DC11, raised using AD brain derived truncated forms of tau proteins, recognizes "Tau DC11 state" that represents all and only those truncated tau proteins that are conformationally different from normal healthy tau proteins (Fig. 1A–C). The effect of these truncated tau proteins on the assembly of microtubules was analyzed. The physiological function of healthy tau is characterized mainly by promotion of microtubule assembly. The tau efficiency in promotion of microtubule assembly can be measured by increase in optical density at 340 nm. DC11 positive truncated tau species, except t99-441, displayed significantly higher microtubule assembly promotion activity than normal healthy tau. Short amino-terminal truncation (t99-441) produces no measurable difference from normal tau. Strikingly N- and C-terminally truncated tau species are promoting robust microtubule assembly, 3–4 times higher (OD_{340}:1.2–1.6) than normal healthy tau (OD_{340}:0.4) (Fig. 1D). For electron microscopy analysis of microtubule assembly was selected mAb DC11 positive double truncated tau species (t151-391) and normal healthy tau (t1-441). Electron micrographs show that normal tau induces formation of thin microtubular networks (Fig. 1E). However, interaction of truncated tau species (t151-391) with tubulin produces abnormally thick microtubular networks (bundles). (Fig. 1F), different in their appearance from normal microtubules under the same magnification (3600×).

3.2. AD-like neurofibrillary pathology induced by truncated tau (t151-391) in vivo

To validate suggested pathological function of truncated tau in vivo, we generated transgenic rat that overexpressed truncated tau (t151-391) in the brain and spinal cord (Fig. 2). The most prominent histopathological feature of transgenic rats was extensive argyrophilic NFT formation (Fig. 3A). No neurofibrillary pathology was found in wild type rats throughout their lifespan. The appearance of NFTs satisfied several histological criteria used to identify neurofibrillary degeneration in the human AD including argyrophilia (Fig. 3A), Congo red birefringence (Fig. 3B) and Thioflavin S reactivity (Fig. 3C).

The load of neurofibrillary pathology was stereologically quantified in the brain stem (gigantocellular reticular nucleus) where the mean NFT density was 690/mm^3 with an observed coefficient of variation of 32.9% (Fig. 3D). The mean NFT:neuron ratio was 1:8 in transgenic animals. Furthermore, immunohistochemical analysis revealed that neurofibrillary tangle formation passed through the histologically well-defined maturation stages. The first stage was characterized by intraneuronal pre-tangles, immunoreactive for phosphorylated tau protein. The antibody AT8 detected the diffuse rod-like phospho-tau accumulations within the cytoplasm. The pre-tangle bearing neurons had detectable nuclei and normal appearance (Fig. 3E). The assembly of pre-tangles resulted in formation of intracellular NFTs in cell bodies (Fig. 3F) and in processes as neuropil threads (NTs). The late developmental stage represented extra-neuronal "ghost" tangles (eNFT) that were present as immunoreactive, densely packed tau fibrils or bundles outside the neurons (Fig. 3G). The cell soma revealed no stainable cytoplasm and nucleus.

3.3. The sarcosyl insoluble tau complexes consisted of human truncated tau and endogenous rat tau protein

To determine whether truncated tau (t151-391) was able to induce maturation of neurofibrillary pathology, manifested

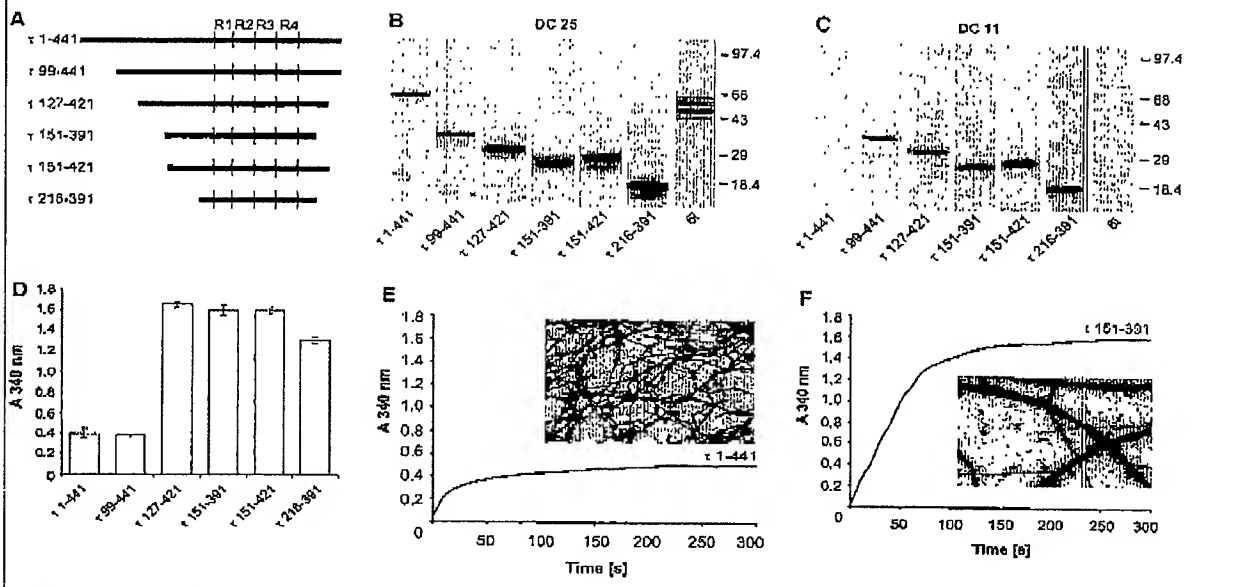


Fig. 1. Efficacy of truncated tau proteins in promotion of microtubule assembly. (A) Schematic diagram of tau species tested in vitro for their potency to promote microtubule assembly. The numbering of amino acids corresponds to that of the human tau 40 [30]. (B, C) Western blot analysis of tau proteins using pan tau mAb DC25 and tau conformation specific mAb DC11. Recombinant human tau six isoforms (6l) were used as a control. Monoclonal antibody DC25 (B) recognizes all tau proteins, however conformation-dependent mAb DC11 (C) stains only truncated forms of tau proteins and does not recognize any of six human tau isoforms. (D) Microtubule assembly induced by truncated tau proteins monitored by turbidometry at OD 340 nm at 5 min. Individual bars reflect efficacy of tau species tested in promotion of microtubule assembly. (E, F) Electron microscopy images of microtubules induced by normal tau (E) and truncated tau (F). Samples were taken at steady state of polymerization (at 5 min). Both figures are at the same magnification (3600×).

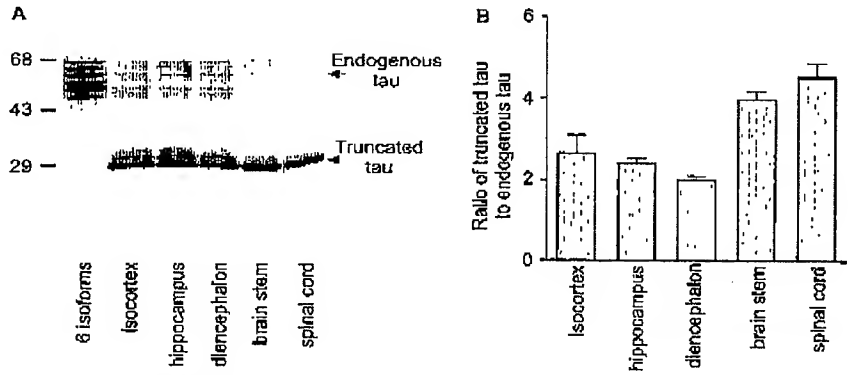


Fig. 2. Expression profile of human truncated tau in the brain and spinal cord of transgenic animals. (A) Pan tau monoclonal antibody DC25 was used for staining of rat endogenous and human truncated tau protein in different brain regions and spinal cord. (B) Transgenic protein expression levels were 2–5-fold over endogenous tau in the isocortex, hippocampus, diencephalon, brain stem and the spinal cord.

by the presence of sarcosyl insoluble tau complexes, we analysed sarcosyl insoluble protein extracts (P2) from 10 to 12 months old transgenic rats, age-matched control rats (wt) and from Alzheimer's diseased brain tissues. Western blot analysis of P2 fraction, from transgenic rat using pan tau mAb DC25, revealed similar staining pattern to that of human AD brain (Fig. 4, lanes 3 and 7). Age-matched control wild type rats had no tau in P2 fraction (Fig. 4, lane 2). To investigate whether human truncated tau (t151-391) co-assembled with endogenous rat tau in transgenic rats, we analyzed the P2 fraction with antibodies reactive with both human and rat tau (DC25), with human tau only (H17) and with endogenous rat tau only (PHF1. human Alzheimer's truncated tau – transgene – does not contain the PHF1 epitope). Our results showed that sarcosyl insoluble P2 fractions from transgenic rats consisted of human tau (Fig. 4, lane 4) and rat endogenous tau (Fig. 4, lane 5). Phosphorylated tau immunoreactivities were detected in the same fraction

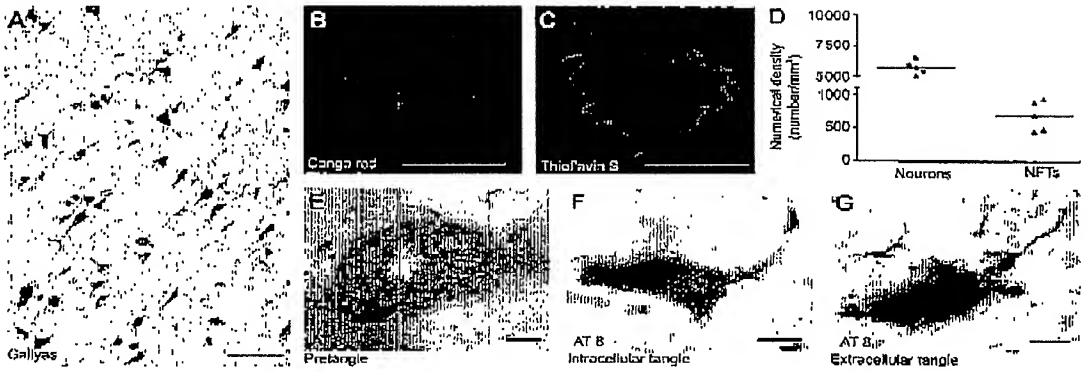


Fig. 3. Truncated tau induced AD-like neurofibrillary degeneration in vivo. (A) Development of extensive argyrophylic positive neurofibrillary tangles in 9 months old rats. High magnification of Congo red (B) and Thioflavin S positive (C) intraneuronal tangles showed similar appearance as in human AD. (D) Stereological analysis of rat brains expressing human truncated tau showed a mean neuronal density of 5703 neurons/mm^3 (S.E.M. = 250.2) in brain stem. The estimated NFT density in this brain region was 690/mm^3 (S.E.M. = 101.4). Ontogeny of the neurofibrillary degeneration in these rats is similar to that of human Alzheimer's disease: pre-tangles (E), intracellular tangles (F) and extracellular tangles (G). Tool bars 10 μm.

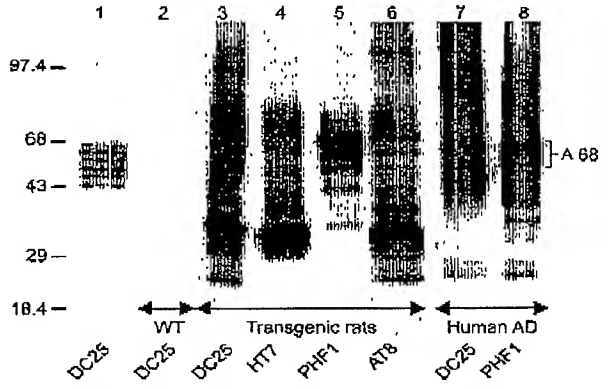


Fig. 4. Human truncated tau and endogenous rat tau are constituent parts of massive sarcosyl insoluble tau complexes. The series of ultracentrifugation and extraction steps was used to obtain sarcosyl insoluble fraction of tau (P2 fraction) from brain tissues of age-matched wild type (wt) rats, transgenic rats and human AD brain tissues. Recombinant human six tau isoforms (lane 1) were used as a control. Wild type rats did not show any sarcosyl insoluble tau (lane 2). Immunoblotting of P2 fraction from transgenic rats revealed sarcosyl insoluble complexes of tau (lane 3, mAb DC25) that were formed from human truncated tau (lane 4, HT7-human tau-specific mAb), endogenous phosphorylated rat tau (lane 5, mAb PHF1-endogenous rat tau specific, see Section 3), and human and rat phosphorylated tau (lane 6, mAb AT8). P2 fraction of human AD is shown as a positive control with characteristic A68 triplet (lane 7, mAb DC25, lane 8, mAb PHF1) seen in rats as well (lane 5).

(P2) with phosphorylation-dependent antibodies, AT8 and PHF1. AT8 phosphosecrines 202 and 205 were present in both human and rat tau (Fig. 4, lane 6). Abnormal phosphorylation of rat endogenous tau was detected by PHF1 that does not recognize human truncated tau (Fig. 4, line 5). Furthermore, it is noteworthy that the A68 triplet characteristic of human AD neurofibrillary degeneration (Fig. 4, lane 8) as revealed by PHF1 staining was found in transgenic animals as well (Fig. 4, lane 5).

3.4. Quantitative analyses of human transgenic and endogenous rat tau in sarcosyl insoluble fraction

We examined further the composition of mature sarcosyl insoluble tau complexes with respect to the ratio between the transgenic human truncated tau (t151-391) and endogenous rat tau. Insoluble P2 fractions from 12 months old rats were assayed on Western blot together with three respective sarcosyl soluble (S2) fractions. Both soluble and insoluble fractions were stained with pan-tau monoclonal antibody DC25 (Fig. 5A) and with human tau-specific monoclonal antibody HT7 (Fig. 5B). Data from immunoblotted transgenic human tau in S2 fractions were digitized and used as a standard for normalization of monoclonal antibodies staining. Both antibodies stained tau protein with similar intensities (Fig. 5E). Tau staining in P2 fraction was digitized as well (Fig. 5C and D) and the relative tau protein amount was calculated on the basis of correlation between amounts of proteins seen by both antibodies. The comparison of normalized peak areas revealed that mature sarcosyl insoluble tau complexes are composed of transgenic human truncated tau and endogenous rat tau at 1:1 ratio (Fig. 5F).

3.5. The level of sarcosyl insoluble formation correlates with lifespan of transgenic rats expressing truncated tau

Sarcosyl insolubility of tau is generally considered to be a definitive transformation point of physiological tau into pathological form. Therefore, we analysed development of sarcosyl insoluble tau complexes in the brain of transgenic rats expressing truncated tau (t151-391). The brain tissues were examined at 3, 6, 9 and 12 months old animals. The level of tau in the sarcosyl insoluble P2 fraction increased in an age-dependent manner and correlated positively with the development of neurofibrillary pathology. First sarcosyl insoluble tau consisting exclusively of transgene – human truncated tau – appeared in the brain of 3 months old transgenic rats and persisted until the late stages of neurodegeneration (Fig. 6A, lanes 1, 2). The first, phosphorylation induced electrophoretic mobility decrease (gel shift)

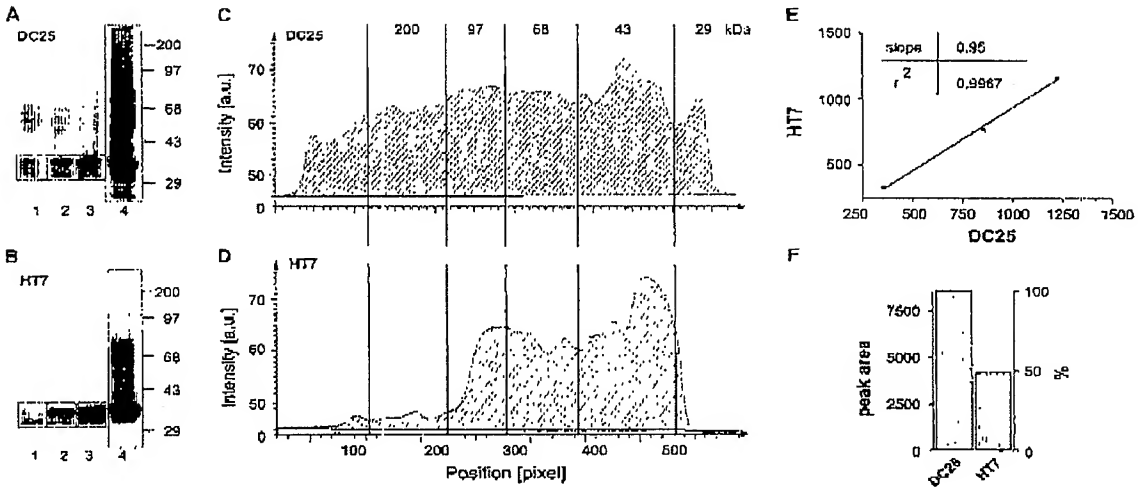


Fig. 5. Quantitative analyses of human transgenic and endogenous rat tau in sarcosyl insoluble fraction. The fractions containing sarcosyl insoluble tau (P2) were analyzed with pan tau monoclonal antibody DC25 and human tau-specific antibody HT7. Lanes 1–3 show sarcosyl soluble fraction (S2) from three independent transgenic animals; lane 4 shows sarcosyl insoluble fraction (P2) from transgenic animal (A, B). Boxed off are P2 fractions (A, B; lane 4) that were used for quantification (C, D). Integrated signals from S2 fractions (A, B; lanes 1–3) were used for construction of the correlation line (E). Ratio between human transgenic and endogenous rat tau in P2 fraction is 1:1 (F). The mAb DC25 staining reflects the total amount of tau present in sarcosyl insoluble P2 pellets, whereas HT7 detects human transgenic tau only.

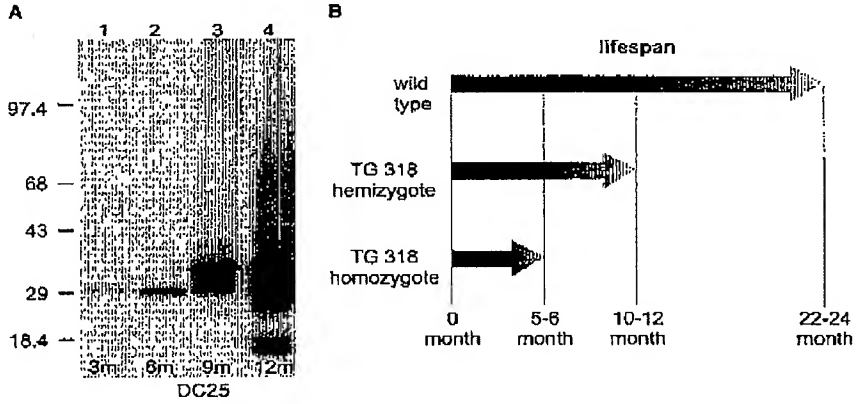


Fig. 6. Ontogenesis of sarcosyl insoluble tau complexes. (A) Ontogenesis of sarcosyl insoluble tau complexes (P2) in the brain of 3, 6, 9 and 12 months old transgenic rats was monitored by Western blot analysis using pan tau mAb DC25. (B) Alzheimer's tau expression impact on lifespan of transgenic rats.

of sarcosyl insoluble tau monomer was observed in 9 months old animals (Fig. 6A, lane 3). Mature sarcosyl insoluble tau formations, characterized by the presence of tau species with high and low molecular weight, appeared in 12 months old animals (Fig. 6A, lane 4). It is noteworthy that the stage of "mature sarcosyl insoluble tau formation" correlated with death of animals expressing transgenic human truncated tau. The lifespan of hemizygous animals was 10–12 months and that of homozygotes was 5–6 months. Life expectation of wild type rats is 22–24 months. These results show that expression of human truncated tau shortens lifespan in hemizygotes by 50% and in homozygotes by 75% (Fig. 6B).

4. Discussion

During the work providing molecular proof that microtubule associated protein tau is a major (if not sole) constituent of paired helical filaments [2,21], it was noted that tau could be truncated. Molecular mapping of the epitope of monoclonal antibody 423, that recognizes tau protein derived from AD brains, revealed for the first time that tau is truncated at E^{391} in Alzheimer's disease [6,12]. Since then, truncation of tau was suggested by many authors as a possible seminal event in the pathogenesis of Alzheimer's disease [22–24]. It is generally agreed that tau has to undergo significant conformational change(s) leading to the pathological polymerization process.

It has been shown that truncation could facilitate polymerization of tau in vitro [25–27]. Despite of these results, the role of truncation and truncated tau in AD cascade remains an open issue. Monoclonal antibody DC11, produced against AD brain derived truncated forms of tau protein, revealed the presence of conformationally distinct forms of tau protein in AD. Molecular analysis of these forms showed that DC11 recognizes all and only those N- and C-terminally truncated tau proteins that are conformationally different from normal healthy tau proteins [8]. The shift of tau into "DC11 state" in AD could represent a new pathogenic entity and important step in neurofibrillary degeneration itself. Therefore, we studied effect of N- and C-terminally truncated tau species, on their capacity to influence microtubule assembly. Strikingly, these double truncated tau species promoted robust microtubule assembly, that was 3–4 times higher (OD_{340}:1.2–1.6) than microtubule assembly induced by normal healthy tau (OD_{340}:0.4). Electron microscopy analysis of microtubules assembled by DC11 tau species revealed abnormally thick microtubular networks (bundles) that differed from normal microtubular networks. These results suggest that the truncated tau has large impact on microtubule assembly in vitro suggesting its possible gain of altered function that could lead to tau transformation into a pathological entity.

In order to explore the possible role of truncated tau in vivo, we expressed the most in vitro active DC11 tau species (t151-391) as a transgene in rat brains. Transgenic animals developed extensive neurofibrillary pathology satisfying several histopathological criteria used for identification of neurofibrillary degeneration in AD, including argyrophilia, Congo red birefringence and Thioflavin S reactivity. As in human AD, formation of NFT in transgenic animals passed through several histologically defined maturation stages. First stage was represented by pre-tangle formation (identified with mAb AT8) that is considered to be an early event in NFT development [28,29]. The second stage was characterized by formation of intracellular argyrophilic NFTs in neuronal cell bodies and NTs in their processes. The late developmental stage in these transgenic animals was represented by the presence of extra-neuronal "ghost" tangles (eNFT). The well-defined staging in transgenic rats expressing truncated tau offers an opportunity to study the neurodegenerative cascade of tau protein in vivo.

In human sporadic AD, mature neurofibrillary degeneration is characterized by extensive formation of sarcosyl insoluble tau protein complexes consisting of abnormally hyperphosphorylated full length and truncated tau forms [1–4]. The analysis of sarcosyl insoluble tau fractions derived from the brain of transgenic animals allowed drawing several important conclusions: First, sarcosyl insoluble tau complexes were composed of transgenic human truncated tau and endogenous rat tau at a 1:1 ratio. Second, both human and endogenous rat tau were phosphorylated (AT8) and third, tau A68 triplet pattern characteristic of human AD [3] was formed in transgenic animals. Furthermore, detailed time course experiments of neurofibrillary maturation revealed that first sarcosyl insoluble truncated tau monomer appeared already in very young transgenic animals (3 months old), well before the detection of intraneuronal tangles (9 months old). We suggest that sarcosyl insoluble monomer ("one band stage") represents immature developmental stage of sarcosyl insoluble complex formations. Further "aging" of sarcosyl insoluble tau is represented by intensive phosphorylation ("stage of shifted monomer"). Most probably the phosphorylated monomers led to the development of mature sarcosyl insoluble tau complexes ("stage of tau ladder") encompassing both truncated and endogenous full-length tau (9–12 months old). It is intriguing that the stage of "mature sarcosyl insoluble tau formation" correlated with the death of animals expressing transgenic human truncated tau. The life span of hemizygous animals was 10–12 months and that of homozygotes was 5–6 months. Life expectation of wild type rats is 22–24 months. Thus truncated tau expression shortens life span of hemizygotes by 50% and of homozygotes by 75%.

The present study provides experimental data introducing truncated tau protein as an important upstream factor in the pathogenesis of neurofibrillary degeneration of AD type. In addition, our data established that truncated tau is sufficient to drive neurofibrillary degeneration in the absence of tau mutation.

References

[1] Grundke-Iqbal, I., Iqbal, K., Tung, Y.C., Quinlan, M., Wisniewski, H.M. and Binder, L.I. (1986) Abnormal phosphorylation of the microtubule-associated protein tau in Alzheimer cytoskeletal pathology. Proc. Natl. Acad. Sci. USA 83, 4913–4917.

[2] Wischik, C.M., Novak, M., Thogersen, H.C., Edwards, P.C., Runswick, M.J., Jakes, R., Walker, J.E., Milstein, C., Roth, M. and Klug, A. (1988) Isolation of a fragment of tau derived from the core of paired helical filament of Alzheimer's disease. Proc. Natl. Acad. Sci. USA 85, 4506–4510.

[3] Lee, V.M., Balin, B.J., Otvos Jr., L. and Trojanowski, J.Q. (1991) A68: a major subunit of paired helical filaments and derivatized forms of normal Tau. Science 251, 675–678.

[4] Goedert, M. and Klug, A. (1999) Tau protein and the paired helical filament of Alzheimer's disease. Brain Res. Bull. 50, 469–470.

[5] Novak, M., Jakes, R., Edwards, P.C., Milstein, C. and Wischik, C.M. (1991) Difference between the tau protein of Alzheimer paired helical filament core and normal tau revealed by epitope analysis of monoclonal antibodies 423 and 7.51. Proc. Natl. Acad. Sci. USA 88, 5837–5841.

[6] Novak, M., Kabat, J. and Wischik, C.M. (1993) Molecular characterization of the minimal protease resistant tau unit of the Alzheimer's disease paired helical filament. EMBO J. 12, 365–370.

[7] Canu, N., Dus, L., Barbato, C., Ciotti, M.T., Brancolini, C., Rinaldi, A.M., Novak, M., Cattaneo, A., Bradbury, A. and Calissano, P. (1998) Tau cleavage and dephosphorylation in cerebellar granule neurons undergoing apoptosis. J. Neurosci. 18, 7061–7074.

[8] Vechterova, L., Kontsekova, E., Zilka, N., Ferencik, M., Ravid, R. and Novak, M. (2003) DC11: a novel monoclonal antibody revealing Alzheimer's disease specific tau epitope. Neuroreport 14, 87–91.

[9] Horowitz, P.M., Patterson, K.R., Guillozet-Bongaarts, A.L., Reynolds, M.R., Carroll, C.A., Weintraub, S.T., Bennett, D.A., Cryns, V.L., Berry, R.W. and Binder, L.I. (2004) Early N-terminal changes and caspase-6 cleavage of tau in Alzheimer's disease. J. Neurosci. 24, 7895–7902.

[10] Guillozet-Bongaarts, A.L., García-Sierra, F., Reynolds, M.R., Horowitz, P.M., Fu, Y., Wang, T., Cahill, M.E., Bigio, E.H., Berry, R.W. and Binder, L.I. (2005) Tau truncation during neurofibrillary tangle evolution in Alzheimer's disease. Neurobiol. Aging 26, 1015–1022.

[11] Novak, M. (1994) Truncated tau protein as a new marker for Alzheimer's disease. Acta Virol. 38, 173–189.

[12] Novak, M., Wischik, C.M., Edwards, P., Pannell, R. and Milstein, C. (1989) Characterisation of the first monoclonal antibody against the pronase resistant core of the Alzheimer PHF. Prog. Clin. Biol. Res. 317, 755–761.

[13] Skrabana, R., Kontsek, P., Mederlyova, A., Iqbal, K. and Novak, M. (2004) Folding of Alzheimer's core PHF subunit revealed by monoclonal antibody 423. FEBS Lett. 568, 178–182.

[14] Csokova, N., Skrabana, R., Liebig, H.D., Mederlyova, A., Kontsek, P. and Novak, M. (2004) Rapid purification of truncated tau proteins: model approach to purification of functionally active fragments of disordered proteins. implication for neurodegenerative diseases. Protein Expr. Purif. 35, 366–372.
[15] Vallee, R.B. (1986) Reversible assembly purification of microtubules without assembly-promoting agents and further purification of tubulin, microtubule-associated proteins, and MAP fragments, Methods Enzymol. 134, 89–104.
[16] Gallyas, F. (1971) Silver staining of Alzheimer's neurofibrillary changes by means of physical development. Acta Morphol. Acad. Sci. Hung. 19, 1–8.
[17] Sun, A., Nguyen, X.V. and Bing, G. (2002) Comparative analysis of an improved Thioflavin-S stain, Gallyas silver stain, and immunohistochemistry for neurofibrillary tangle demonstration on the same sections. J. Histochem. Cytochem. 50, 463–472.
[18] Gundersen, H.J., Bagger, P., Bendtsen, T.F., Evans, S.M., Korbo, L., Marcussen, N., Moller, A., Nielsen, K., Nyengaard, J.R. and Pakkenberg, B. (1988) The new stereological tools: disector, fractionator, nucleator and point sampled intercepts and their use in pathological research and diagnosis. Apmis 96, 857–881.
[19] Dorph-Petersen, K.A., Nyengaard, J.R. and Gundersen, H.J. (2001) Tissue shrinkage and unbiased stereological estimation of particle number and size. J. Microsc. 204, 232–246.
[20] Greenberg, S.G. and Davies, P. (1990) A preparation of Alzheimer paired helical filaments that displays distinct τ proteins by polyacrylamide gel electrophoresis. Proc. Natl. Acad. Sci. USA 87, 5827–5831.
[21] Wischik, C.M., Novak, M., Edwards, P.C., Klug, A., Tichelaar, W. and Crowther, R.A. (1988) Structural characterization of the core of the paired helical filament of Alzheimer disease. Proc. Natl. Acad. Sci. USA 85, 4884–4888.
[22] Ghoshal, N., Garcia-Sierra, F., Fu, Y., Beckett, L.A., Mufson, E.J., Kuret, J., Berry, R.W. and Binder, L.I. (2001) Tau-66: evidence for a novel tau conformation in Alzheimer's disease. J. Neurochem. 77, 1372–1385.
[23] Garcia-Sierra, F., Ghoshal, N., Quinn, B., Berry, R.W. and Binder, L.I. (2003) Conformational changes and truncation of tau protein during tangle evolution in Alzheimer's disease. J. Alzheimers Dis. 5, 65–77.
[24] Binder, L.I., Guillozet-Bongaarts, A.L., Garcia-Sierra, F. and Berry, R.W. (2005) Tau, tangles, and Alzheimer's disease. Biochim. Biophys. Acta 1739, 216–223.
[25] Abraha, A., Ghoshal, N., Gamblin, T.C., Cryns, V., Berry, R.W., Kuret, J. and Binder, L.I. (2000) C-terminal inhibition of tau assembly in vitro and in Alzheimer's disease. J. Cell Sci. 21, 3737–3745.
[26] Iqbal, K., Alonso, C., Chen, S., Chohan, M.O., El-Akkad, E., Gong, C.X., Khatoon, S., Li, B., Liu, F., Rahman, A., Tanimukai, H. and Grundke-Iqbal, I. (2005) Tau pathology in Alzheimer disease and other tauopathies. Biochim. Biophys. Acta 1739, 198–210.
[27] Avila, J. (2006) Tau phosphorylation and aggregation in Alzheimer's disease pathology. FEBS Lett. 580, 2922–2927.
[28] Braak, E., Braak, H. and Mandelkow, E.M. (1994) A sequence of cytoskeleton changes related to the formation of neurofibrillary tangles and neuropil threads. Acta Neuropathol. 87, 554–567.
[29] Augustinack, J.C., Schneider, A., Mandelkow, E.M. and Hyman, B.T. (2002) Specific tau phosphorylation sites correlate with severity of neuronal cytopathology in Alzheimer's disease. Acta Neuropathol. 103, 26–35.
[30] Goedert, M., Spillantini, M.G., Jakes, R., Rutherford, D. and Crowther, R.A. (1989) Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. Neuron 3, 519–526.

EXHIBIT 3

Figure 1

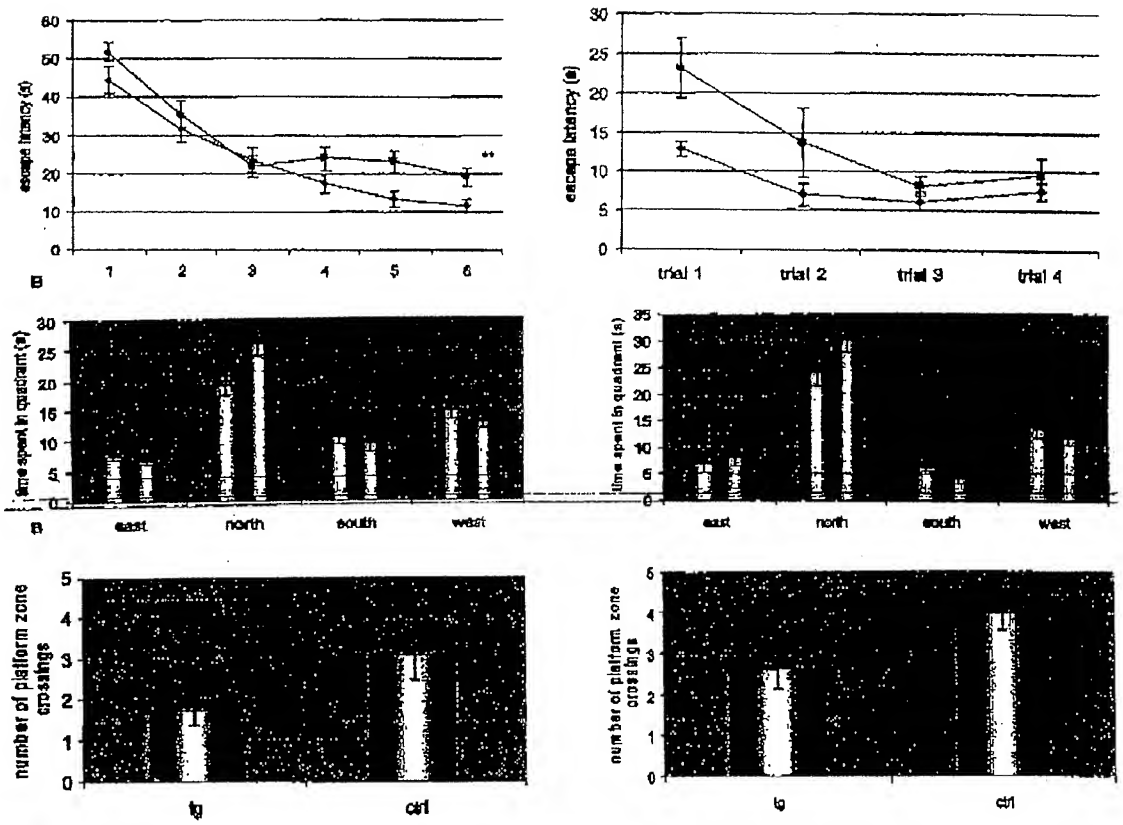
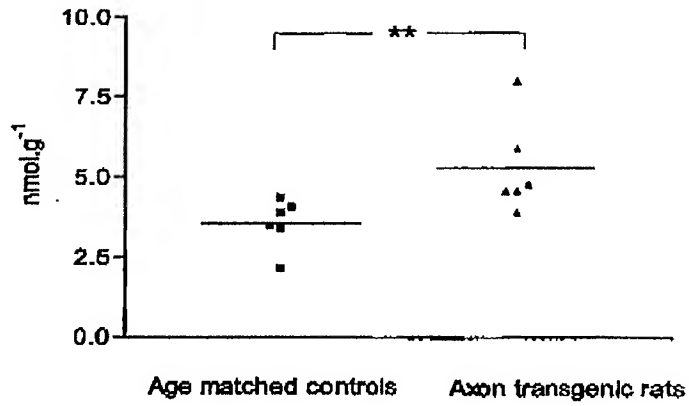


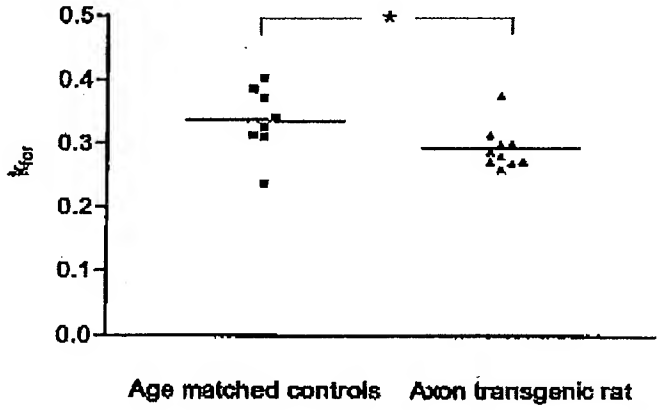
Fig.1A. Statistically significant difference was observed in the acquisition of the spatial information in transgenic rats relative to their non-transgenic litter-mates at the age of 5 months. The escape latency (A) to find the hidden platform over the 6 days (4 daily trials) of the training phase (RM-ANOVA, ** $P<0.01$).
Fig.1B. Water maze visible platform acquisition during 4 testing trials over one day. Visible platform showed no loss of motivation or visual acuity in tested rats.
Fig.1C, D. Significant difference in time spent in target quadrant (north) between transgenic rats compared to controls (t-test, *$P<0.05$) during the probe trial measured after three days of acquisition learning (Fig.1C). The difference in the probe trial performed after six days (Fig.1D) did not reach statistical significance (t-test, $P=0.1$).
Fig.1E, F. Number of crosses over the platform location was significantly lower in transgenic rats if compared to wild type controls (t- test, *$P<0.05$) during the first probe trial (Fig.1E). The number of platform crosses during the second probe trial (Fig.1F) did not reach significance (t- test, *$P=0.06$). Values represent mean ±S.E.M.

Figure 2



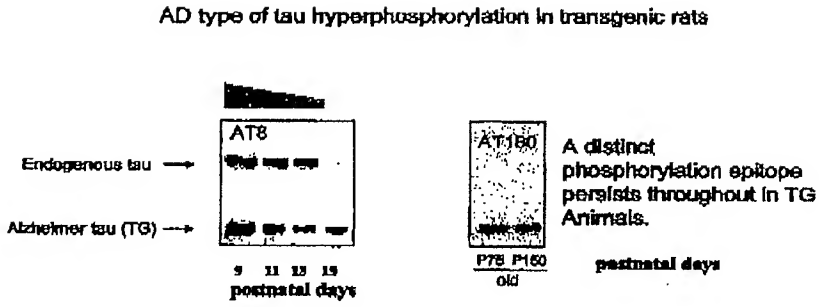
The ascorbate free radical electron paramagnetic resonance (EPR) signal show higher concentration of AFR in the homogenate obtained from brain stems of animals (5.300 nmol.g-1 ± 0.6011, N=6) than from age-matched control rats (3.583 nmol.g-1 ± 0.3156 N=6). Increased amount of AFR (P<0.01) in the brain of transgenic rats at the terminal stage indicate, that oxidative stress is a consequence and not a cause of pathological cascade in the transgenic rats.

Figure 3



The kinetic measurement of CK reaction showed, that the rate constant values of the brain specific creatine kinase (CKBB) was significantly decreased in the brain of Axon transgenic rats. The kinetic measurement of CK reaction showed, that the rate constant values of the CKBB was significantly decreased (P<0.05) in the brain of Axon transgenic rats (k for = 0.2942 ± 0.01048, N=10) in comparison with age-matched controls (k for = 0.3370 ± 0.01862, N=8).

Figure 4



Using mAB AT180 (Fig.4, right panel), strong AD-like phosphorylation persists in brain extracts from 75 and 150 day old animals. This type of phosphorylation (right panel) drives further development of neurofibrillary changes identical to human AD. None of these features is seen any previous tau transgenic animal using wild type or FTDP17 mutated tau transgene construct. The left panel of Fig. 4 shows the typical phosphorylation pattern of endogenous tau in embryos/early days p.n. which drops by day 19 p.n. In contrast, transgenic truncated tau remains phosphorylated thus reflecting distinct mechanisms of tau phosphorylation.

EXHIBIT 4

Genome sequence of the Brown Norway rat yields insights into mammalian evolution

Rat Genome Sequencing Project Consortium*

**Lists of participants and affiliations appear at the end of the paper*

The laboratory rat (*Rattus norvegicus*) is an indispensable tool in experimental medicine and drug development, having made inestimable contributions to human health. We report here the genome sequence of the Brown Norway (BN) rat strain. The sequence represents a high-quality 'draft' covering over 90% of the genome. The BN rat sequence is the third complete mammalian genome to be deciphered, and three-way comparisons with the human and mouse genomes resolve details of mammalian evolution. This first comprehensive analysis includes genes and proteins and their relation to human disease, repeated sequences, comparative genome-wide studies of mammalian orthologous chromosomal regions and rearrangement breakpoints, reconstruction of ancestral karyotypes and the events leading to existing species, rates of variation, and lineage-specific and lineage-independent evolutionary events such as expansion of gene families, orthology relations and protein evolution.

Darwin believed that "natural selection will always act very slowly, often only at long intervals of time"[1]. The consequences of evolution over timescales of approximately 1,000 millions of years (Myr) and 75 Myr were investigated in publications comparing the human with invertebrate and mouse genomes, respectively[2,3]. Here we describe changes in mammalian genomes that occurred in a shorter time interval, approximately 12–24 Myr (refs 4, 5) since the common ancestor of rat and mouse.

The comparison of these genomes has produced a number of insights:

- The rat genome (2.75 gigabases, Gb) is smaller than the human (2.9 Gb) but appears larger than the mouse (initially 2.5 Gb (ref. 3) but given as 2.6 Gb in NCBI build 32, see http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html).
- The rat, mouse and human genomes encode similar numbers of genes. The majority have persisted without deletion or duplication since the last common ancestor. Intronic structures are well conserved.
- Some genes found in rat, but not mouse, arose through expansion of gene families. These include genes producing pheromones, or involved in immunity, chemosensation, detoxification or proteolysis.
- Almost all human genes known to be associated with disease have orthologues in the rat genome but their rates of synonymous substitution are significantly different from the remaining genes.
- About 3% of the rat genome is in large segmental duplications, a fraction intermediate between mouse (1–2%) and human (5–6%). These occur predominantly in pericentromeric regions. Recent expansions of major gene families are due to these genomic duplications.
- The eutherian core of the rat genome—that is, bases that align orthologously to mouse and human—comprises a billion nucleotides (~40% of the euchromatic rat genome) and contains the vast majority of exons and known regulatory elements (1–2% of the genome). A portion of this core constituting 5–6% of the genome appears to be under selective constraint in rodents and primates, while the remainder appears to be evolving neutrally.
- Approximately 30% of the rat genome aligns only with mouse, a considerable portion of which is rodent-specific repeats. Of the non-aligning portion, at least half is rat-specific repeats.
- More genomic changes occurred in the rodent lineages than the primate: (1) These rodent genomic changes include approximately 250 large rearrangements between a hypothetical murid ancestor and human, approximately 50 from the murid ancestor to rat, and about the same from the murid ancestor to mouse. (2) A threefold-higher rate of base substitution in neutral DNA is found along the rodent lineage when compared with the human lineage, with the rate on the rat branch 5–10% higher than along the mouse branch. (3) Microdeletions occur at an approximately twofold-higher rate than microinsertions in both rat and mouse branches.
- A strong correlation exists between local rates of microinsertions and microdeletions, transposable element insertion, and nucleotide substitutions since divergence of rat and mouse, even though these events occurred independently in the two lineages.

Background

History of the rat

The rat, hated and loved at once, is both scourge and servant to mankind. The "Devil's Lapdog" is the first sign in the Chinese zodiac and traditionally carries the Hindu god Ganesh[6]. Rats are a reservoir of pathogens, known to carry over 70 diseases. They are involved in the transmission of infectious diseases to man, including cholera, bubonic plague, typhus, leptospirosis, cowpox and hantavirus infections. The rat remains a major pest, contributing to famine with other rodents by eating around one-fifth of the world's food harvest.

Paradoxically, the rat's contribution to human health cannot be overestimated, from testing new drugs, to understanding essential nutrients, to increasing knowledge of the pathobiology of human disease. In many parts of the world the rat remains a source of meat.

The laboratory rat (*R. norvegicus*) originated in central Asia and its success at spreading throughout the world can be directly attributed to its relationship with humans[7]. J. Berkenhout, in his 1769 treatise *Outline of the Natural History of Great Britain*, mistakenly took it to be from Norway and used *R. norvegicus* Berkenhout in the first formal Linnaean description of the species. Whereas the black rat (*Rattus rattus*) was part of the European landscape from at least the third century AD and is the species associated with the spread of bubonic plague, *R. norvegicus* probably originated in northern China and migrated to Europe somewhere

around the eighteenth century[8]. They may have entered Europe after an earthquake in 1727 by swimming the Volga river.

The rat in research

R. norvegicus was the first mammalian species to be domesticated for scientific research, with work dating to before 1828 (ref. 9). The first recorded breeding colony for rats was established in 1856 (ref. 9). Rat genetics had a surprisingly early start. The first studies by Crampe from 1877 to 1885 focused on the inheritance of coat colour[10]. Following the rediscovery of Mendel's laws at the turn of the century, Bateson used these concepts in 1903 to demonstrate that rat coat colour is a mendelian trait[10]. The first inbred rat strain, PA, was established by King in 1909, the same year that systematic inbreeding began for the mouse[10]. Despite this, the mouse became the dominant model for mammalian geneticists, while the rat became the model of choice for physiologists, nutritionists and other biomedical researchers. Nevertheless, there are over 234 inbred strains of *R. norvegicus* developed by selective breeding, which 'fixes' natural disease alleles in particular strains or colonies[11].

Over the past century, the role of the rat in medicine has transformed from carrier of contagious diseases to indispensable tool in experimental medicine and drug development. Current examples of use of the rat in human medical research include surgery[12], transplantation[13–15], cancer[16,17], diabetes[18,19], psychiatric disorders[20] including behavioural intervention[21] and addiction[22], neural regeneration[23,24], wound[25,26] and bone healing[27], space motion sickness[28], and cardiovascular disease[29–31]. In drug development, the rat is routinely employed both to demonstrate therapeutic efficacy[15,32,33] and to assess toxicity of novel therapeutic compounds before human clinical trials[34–37].

The Rat Genome Project

Over the past decade, investigators and funding agencies have participated in rat genomics to develop valuable resources. Before the launch of the Rat Genome Sequencing Project (RGSP), there was much debate about the overall value of the rat genome sequence and its contribution to the utility of the rat as a model organism. The debate was fuelled by the naive belief that the rat and mouse were so similar morphologically and evolutionarily that the rat sequence would be redundant. Nevertheless, an effort spearheaded by two NIH agencies (NHGRI and NHLBI) culminated in the formation of the RGSP Consortium (RGSPC).

The RGSP was to generate a draft sequence of the rat genome, and, unlike the comparable human and mouse projects, errors would not ultimately be corrected in a finished sequence[38]. Consequently, the draft quality was critical. Although it was expected to have gaps and areas of inaccuracy, the overall sequence quality had to be high enough to support detailed analyses.

The BN rat was selected as a sequencing target by the research community. An inbred animal (BN/SsNHsd) was obtained by the Medical College of Wisconsin (MCW) from Harlan Sprague Dawley. Microsatellite studies indicated heterozygosity, so over 13 generations of additional inbreeding were performed at the MCW, resulting in BN/SsNHsd/Mcwi animals. Most of the sequence data were from two females, with a small amount of whole genome shotgun (WGS) and flow-sorted Y chromosome sequencing from a male. The Y chromosome is not included in the current assembly.

A network of centres generated data and resources, led by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) and including Celera Genomics, the Genome Therapeutics Corporation, the British Columbia Cancer Agency Genome Sciences Centre, The Institute for Genomic Research, the University of Utah, the Medical College of Wisconsin, The Children's Hospital of Oakland Research Institute, and the Max Delbrück Center for Molecular Medicine, Berlin. After assembly of the genome at the BCM-HGSC, analysis was performed by an international team, representing over 20 groups in six countries and relying largely on gene and protein predictions produced by Ensembl.

Determination of the genome sequence

Atlas and the 'combined' sequencing strategy

Despite progress in assembling draft sequences[2,3,39–44] the question of which method produces the highest-quality products is unresolved. A significant issue is the choice between logistically simpler WGS approaches versus more complex strategies employing bacterial artificial chromosome (BAC) clones[45–48]. In the Public Human Genome Project[2] a BAC by BAC hierarchical approach was used and provided advantages in assembling difficult parts of the genome. The draft mouse sequence was a pure WGS approach using the ARACHNE assembler[3,49,50] but underrepresented duplicated regions owing to 'collapses' in the assembly[3,51–53]. This limitation of the mouse draft sequence was tolerable owing to the planned full use of BAC clones in constructing the final finished sequence.

The RGSPC opted to develop a 'combined' approach using both WGS and BAC sequencing (Fig. 1). In the combined approach, WGS data are progressively melded with light sequence coverage of individual BACs (BAC skims) to yield intermediate products called 'enriched BACs' (eBACs). eBACs covering the whole genome are then joined into longer structures (bactigs). Bactigs are joined to form larger structures: superbactigs, then ultrabactigs. During this process other data are introduced, including BAC end sequences, DNA fingerprints and other long-range information (genetic markers, syntenic information), but the process is constrained by eBAC structures.

To execute the combined strategy we developed the *Atlas* software package[54] (Fig. 1). The *Atlas* suite includes a 'BAC-Fisher' component that performs the functions needed to generate eBACs. WGS genome coverage was generated ahead of complete BAC coverage, so a BAC-Fisher web server was established at the BCM-HGSC to enable users to access the combined BAC and WGS reads as each BAC was processed (see Methods for data access). Each eBAC is assembled with high stringency to represent the local sequence accurately, and so provide a valuable intermediate product that assists all users of the genome data. Additional *Atlas* modules joined eBACs and linked bactigs to give the complete assembly (Fig. 1). Overall, the combined approach takes advantage of the strengths of both previous methods, with few of the disadvantages.

Sequence and genome data

Over 44 million DNA sequence reads were generated (Table 1; Methods). Following removal of low-quality reads and vector contaminants, 36 million reads were used for *Atlas* assembly, which retained 34 million reads. This was 7× sequence coverage with 60% provided by WGS and 40% from BACs. Slightly different estimates came from considering the entire 'trimmed' length of the sequence data (7.3×), or only the portion of Phred20 quality or higher (6.9×).

The sequence data were end-reads from clones either derived directly from the genome (insert sizes of <10 kb, 10 kb, 50 kb and >150 kb) or from small insert plasmids subcloned from BACs. Overall, these provided 42-fold clone coverage, with 32-fold coverage having both paired ends represented. Approximately equal contributions of clone coverage were from the different categories.

Over 21,000 BACs were used for BAC skims (1.6× coverage) with an average sequence depth of 1.8×, giving an overall 2.8× genomic sequence coverage from BACs. This was slightly more than the most efficient procedure would require (~1.2× each), because the genome size was not known at the project start.

Simultaneous with sequencing, 199,782 clones from the CHORI-230 BAC library[55] were fingerprinted by restriction enzyme

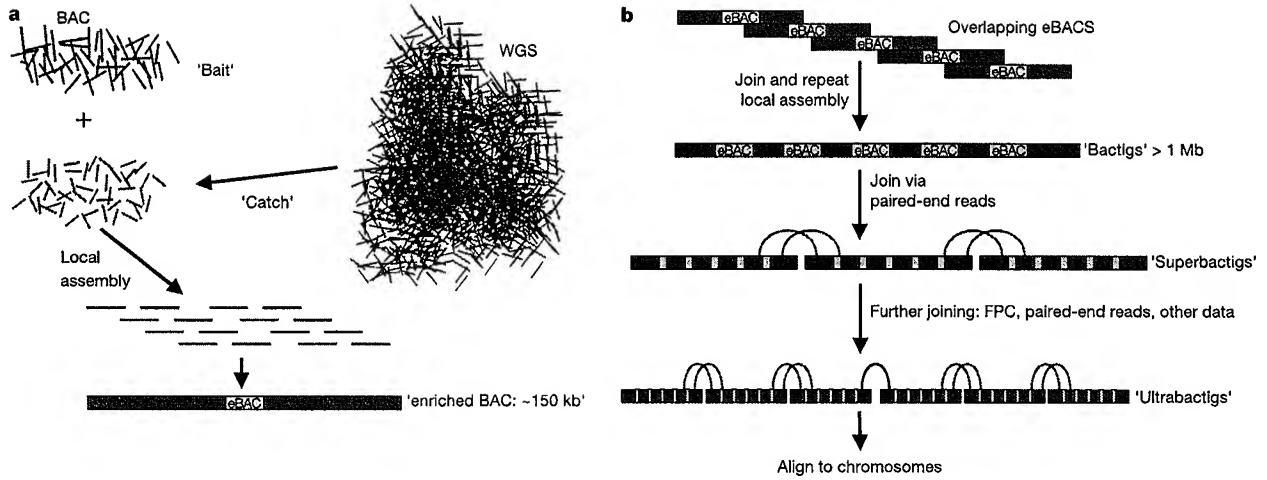


Figure 1 The new 'combined' sequence strategy and *Atlas* software. **a**, Formation of 'eBACs'. The RGSP strategy combined the advantages of both BAC and WGS sequence data[54]. Modest sequence coverage (~1.8-fold) from a BAC is used as 'bait' to 'catch' WGS reads from the same region of the genome. These reads, and their mate pairs, are assembled using Phrap to form an eBAC. This stringent local assembly retains 95% of the 'catch'. **b**, Creation of higher-order structures. Multiple eBACs are assembled into bactigs based on sequence overlaps. The bactigs are joined into superbactigs by large clone mate-pair information (at least two links), extended into ultrabactigs using additional information (single links, FPC contigs, synteny, markers), and ultimately aligned to genome mapping data (radiation hybrid and physical maps) to form the complete assembly.

digestion, representing 12-fold genomic coverage[56] (Methods). These were assembled into a 'fingerprint contig (FPC)' map (a contig is a set of overlapping segments of DNA) containing 11,274 FPCs. BAC selection for sequence skimming was based on overlaps between BACs using FPC mapping[56] (M.K. and C.F., unpublished work), ongoing BAC end sequencing (S.Z., unpublished work), and BAC sequence skimming[57]. This strategy led to the sequence of a tiling path of BAC clones, covering the whole genome. In addition to the FPC map, a yeast artificial chromosome (YAC)-based physical map was constructed. 5,803 BAC and P1-derived artificial chromosome (PAC) clones from RPCI-32 and RPCI-31 libraries[55], respectively, were anchored to 51,323 YAC clones originating from two tenfold-coverage YAC libraries[58,225] assembled into 605 contigs[56]. This map was subsequently integrated with the FPC map and the sequence assembly, reducing the total number of map contigs to 376 (minimum length of contig containing the 'typical' nucleotide, N_{50} = 172 clones, 4.4 Mb; 358 anchored to the sequence assembly; Supplementary Information).

The combined strategy enabled development of resources such as the FPC map, BAC end sequences, and BAC skim sequences in parallel, rather than sequentially. In addition to allowing ongoing quality checking, this permitted the data-gathering phase of the project to be completed in less than two years.

Atlas assembly

Statistics for the Rnor3.1 assembly are in Table 2. Contigs within eBACs were ordered and oriented using read-pair information. Read-pair information was also used to add WGS reads to eBACs, even when sequence overlaps could not be reliably detected owing to repeated sequences. BAC skim reads with repeats were included in the assembly of eBACs because they clearly originated within BAC insert sequences. Over 19,000 eBACs were eventually generated.

More than 98% of eBACs were successfully merged to form bactigs (Fig. 1). Bactigs were subsequently reassembled to process all reads from overlapping BACs simultaneously, and then ordered and oriented with respect to each other using FPC map and BAC end sequence read-pair information. These superbactig and ultrabactig structures (see below) were aligned with chromosomes using external information, such as positions of genetic markers. Ultrabactigs represented the largest sequence units used to build chromosomes.

The current release of the rat genome assembly, version Rnor3.1,

Table 1 **Clones and reads used in the RGSP**

Insert size* (kb)	Source or vector	Reads (millions)				Bases (billions)		Sequence coverage†		Clone coverage‡
		All§	Used	Paired	Assembled	Trimmed	≥Phred20	Trimmed	≥Phred20	
2–4	Plasmid	9.6	8.6	7.4	7.9	4.8	4.5	1.8	1.6	3.70
4.5–7.5	Plasmid	4.5	4.3	3.6	3.6	2.4	2.3	0.87	0.82	2.96
10	Plasmid	8.4	7.2	6.4	6.4	4.1	3.8	1.5	1.4	11.63
50	Plasmid	1.7	1.3	1.0	1.1	0.69	0.65	0.25	0.24	9.47
150–250	BAC	0.32	0.31	0.26	0.26	0.18	0.16	0.07	0.06	9.26
Total WGS		24.5	21.7	18.7	19.2	12.1	11.3	4.4	4.1	37.0
2–5	BAC skims	19.6	14.6	13.2	14.5	8.0	7.7	2.9	2.8	4.8‖
Total		44.1	36.3	31.9	33.7	20.2	19.0	7.3	6.9	41.8

*Grouped in ranges of sizes for individual libraries tracked to specific multiples of 0.5 kb.
†Total bases in used reads divided by sampled genome size including all cloned and sequenced euchromatic or heterochromatic regions.
‡Estimated as sum of insert sizes divided by sampled genome size.
§WGS reads available on the NCBI Trace Archive as of 21 March 2003; BAC skim reads attempted at BCM-HGSC as of 12 May 2003; BAC end reads obtained directly from TIGR.
‖Refers to coverage from 2–5 kb subclones from BACs. The BACs that were skimmed amounted to 1.58 × clone coverage.

Table 2 **Statistics of the RGSP draft sequence assembly**

Features*	Number	N50 length (kb)	Bases (Gb)	Bases plus gaps† (Gb)	Percentage of genome‡			
					Sampled (2.78 Gb)		Assembled (2.75 Gb)	
					Bases	Bases + gaps	Bases	Bases + gaps
Anchored contigs	127,810	38	2.476	2.481	89.1	89.2	90.0	90.2
Anchored superbactig scaffolds	783	5,402	2.476	2.509	89.1	90.3	90.0	91.2
Anchored ultrabactigs	291	18,985	2.476	2.687	89.1	96.6	90.0	97.7
Unanchored superbactigs, main scaffolds	134	1,210	0.056	0.062	2.0	2.2	2.0	2.3
Unanchored ultrabactigs	128	1,529	0.056	0.069	2.0	2.5	2.0	2.5
All superbactigs, main scaffolds	917	5,301	2.533	2.571	91.1	92.5	92.1	93.5
Minor scaffolds	4,345	8	0.033	0.038	1.2	1.4	1.2	1.4

*Anchored sequences are those that can be placed on chromosomes because they contain known markers. The main scaffold for each superbactig is the largest set of contigs (in terms of total contig sequence) that can be ordered and oriented using mate-pair links and ordering of BACs. Scaffolds that cannot be ordered and oriented with respect to the main scaffold are termed minor scaffolds.
†Ambiguous bases (N) are counted in the gap sizes, and excluded in the base counts.
‡Computed as bases plus gaps divided by estimated genome size. Sampled genome size is based on oligonucleotide frequency statistics of unassembled WGS reads. Assembled genome size is based on cumulative contig sequence following assembly.

was generated using the data in Table 1. Earlier releases (Rnor2.0/2.1, Methods) were used for a substantial part of the annotation and analysis of genes and proteins, whereas the current release provided the genome description. Rnor3.1 has 128,000 contigs, with N_{50} length 38 kb—larger than the expected genomic extent of a mammalian gene. These sequence contigs were linked into 783 superbactigs that were anchored to the radiation hybrid map[59]. These larger units had N_{50} length 5.4 Mb. Another 134 smaller superbactigs (N_{50} length 1.2 Mb) could not be anchored, presumably because they fell into gaps between markers or because they were in repeated regions that could not be unambiguously placed. From placement on the radiation hybrid map, adjacent superbactigs were further linked to maximize continuity of sequence if appropriate read-pair mates existed or FPC suggested links. This reduced linked superbactigs to 419 pieces with 71 singletons. 291 ultrabactigs with N_{50} length of nearly 19 Mb were placed on chromosomes. Orthology information with mouse and human sequences was also used to resolve conflicts and suggest placement of sequence units. Most of the 128 unplaced units were either singletons or small superbactigs that consisted of few clones. Thus, nearly the entire genome was represented in less than 300 large sequence units.

Quality assessment

Thirteen megabases of high-quality finished rat sequence from BACs were available for comparison with Rnor3.1 (Methods). This analysis showed that the majority of draft bases from within contigs were high quality (1.32 mismatches per 10 kb). This is essentially the accepted accuracy standard for finished sequence (1.0 errors per 10 kb)[60], so the overwhelming majority of contig bases are highly accurate. The highest frequency of mismatches occurred at the ends of contigs. We calculate the average size of these lower-accuracy regions to be 750 base pairs (bp) and they amount to less than 0.9% of the genome. These regions arise from misassembly of terminal reads due to repeated sequences.

Few mismatches were found within contigs. Six were found within contigs when compared with the 13 Mb of finished sequence, or one case per 2.2 Mb. All were insertions or deletions and may represent polymorphisms. Thus, at the fine structure level, the bulk of sequences that make up contigs is nearly the quality of finished sequence.

We judged accuracy of assembly at the chromosomal level by alignment with linkage maps[61] and radiation hybrid map[59] (Fig. 2). Thirteen markers out of 3,824 from the SHRSP × BN genetic map were placed on different chromosomes in the assembly and in the genetic map. Similarly, of the 20,490 sequence tagged sites placed on both the assembly and radiation hybrid (v3.4) map, 96.9% had consistent chromosome placement[59]. Initial alignments identified regions of misassembly, and these were corrected, so that in Rnor3.1 the maps are congruent except for possible mismapped markers. The distribution of assembled sequence among the chromosomes and chromosome sizes in Rnor3.1 are in Supplementary Table SI-2.

Landscape and evolution of the rat genome

Genome size

Genomic assemblies are usually smaller than the actual genome size owing to under-representation of sequences affected by cloning bias, and sequencing and assembly difficulties. Simply equating the assembled genome size with the euchromatic, cloneable portion does not take into account heterochromatin that may be included[62]. We therefore estimated both an assembled genome size, scaled by the inverse of the fraction of features (genetic markers, expressed sequence tags (ESTs), and so on) found in the Rnor3.1 assembly, and a cloneable (or sampled) genome size, which was the part of the genome present in the WGS reads before assembly, as measured by analysing the distribution of short oligomers[63]. The former may be an underestimate because non-repetitive, easily assembled regions can be enriched for known features. The latter should be an overestimate because there are likely to be regions (such as repeats) that can be cloned and sequenced, but not assembled.

For the rat genome, the assembled and cloneable genome sizes are very close. Considering the fraction of the marker set successfully mapped to Rnor3.1 (92%), or the fraction of sequence finished outside the BCM-HGSC (to reduce bias) present in Rnor3.1 (91%), together with the assembled bases in main scaffolds (2.533 Gb, Table 2), we suggest a genome size of 2.75 Gb. Alternatively, analysis of the WGS oligomers of length 24 to 32 predicted a genome size of between 2.76 and 2.81 billion bases. We have used the more conservative value of 2.75 Gb for the rat genome size, but this is

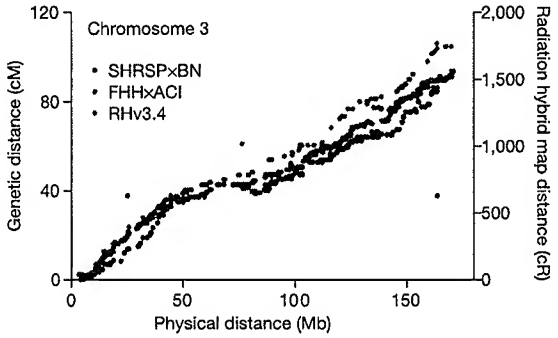


Figure 2 Map correspondence. Correspondence between positions of markers on two genetic maps of the rat (SHRSP × BN intercross and FHH × ACI intercross[61]), on the rat radiation hybrid map[59], and their position on the rat genome assembly (Rnor3.1).

still considerably higher (150 Mb) than the 2.6 Gb currently reported for the mouse draft genome sequence. A fraction of the size differences in these rodent genomes results from the different repeat content (see below); however, it is also recognized that segmental duplications may be under-represented in the mouse WGS draft sequence for technical reasons[3,51].

Telomeres, centromeres and mitochondrial sequence

The rat has both metacentric and telocentric chromosomes, in contrast to the wholly telocentric mouse chromosomes. As expected from previous draft sequences, the rat draft does not contain complete telomeres or centromeres. Their physical location relative to the rat draft sequence can however be approximated; the centromeres of the telocentric rat chromosomes (2, 4–10 and X) must be positioned before nucleotide 1 of these assemblies, and those for the remaining chromosomes are estimated as indicated in Fig. 3. Several of these putative centromere positions coincide with both segmental duplication blocks (see below) and classical satellite clusters, consistent with enrichment of both of these sequence features in rat pericentromeric DNA. Human subtelomere regions are characterized by both an abundance of segmentally duplicated DNA and an enrichment of internal $(TTAGGG)_n$-like sequence islands[64]. Approximately one-third of the euchromatic rat subtelomeric regions are similarly enriched, suggesting that Rnor3.1 might extend very close to the chromosome ends.

Fragments of the rat mitochondrial genome were also propagated within the WGS libraries and subsequently sequenced, allowing the assembly of the complete 16,313 bp mitochondrial genome (Supplementary Information). Comparison with existing mitochondrial sequences in the public databases revealed variable positions totalling 95 bp (0.6%) between this strain and the wild brown rat. Considerably more variation (2.2%) was found when compared with the Wistar strain: 357 bp differences over the whole genome, including 78 positions that are conserved in the other mammalian sequences. Such variation has also been reported in mouse mitochondrial sequences and attributed to errors in previously

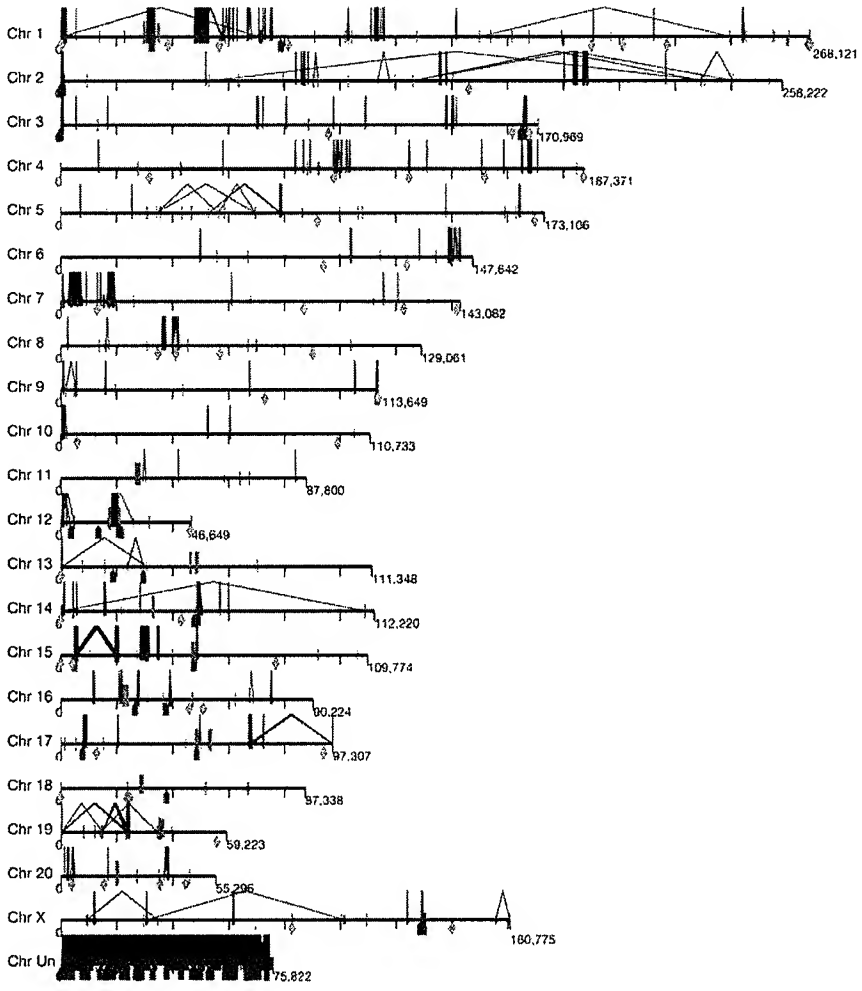


Figure 3 Distribution of segmental duplications in the rat genome. Interchromosomal duplications (red) and intrachromosomal duplications (blue) are depicted for all duplications with ≥90% sequence identity and ≥20 kb length. The intrachromosomal duplications are drawn with connecting blue line segments; those with no apparent connectors are local duplications (spaced below the figure resolution limit). p arms are on the left and the q arms on the right. Chromosomes 2, 4–10, and X are telocentric; the assemblies begin with pericentric sequences of the q arms, and no centromeres are indicated. For the remaining chromosomes, the approximate centromere positions were estimated from the most proximal STS/gene marker to the p and q arm as determined by fluorescent *in situ* hybridization (FISH) (cyan vertical lines; no chromosome 3 data). The 'Chr Un' sequence consists of contigs not incorporated into any chromosomes. Green arrows indicate 1 Mb intervals with more than tenfold enrichment of classic rat satellite repeats within the assembly. Orange diamonds indicate 1 Mb intervals with more than tenfold enrichment of internal $(TTAGGG)_n$-like sequences. For more detail see http://ratparalogy.cwru.edu.

sequenced genomes[65]. The current sequence is very accurate, and we therefore favour the BN sequence as a reference for the rat mitochondrial genome.

Orthologous chromosomal segments and large-scale rearrangements

Multi-megabase segments of the chromosomes of the primate–rodent ancestor have been passed on to human and murid rodent descendants with minimal rearrangements of gene order[66–68]. These intact regions, which are bounded by the breaks that occurred during ancient large-scale chromosomal rearrangements, are referred to as orthologous chromosomal segments. The same phenomenon has occurred in the descent of the rat and mouse from the genome of their common murid ancestor, and we were able to use the human genome, and in some cases other outgroup data, to tentatively reconstruct the sequence of many of these rearrangements in these lineages. To visualize the extent of orthologous chromosomal segments, each genome was 'painted' with the orthologous segments of the other two species (Fig. 4) using the Virtual Genome Painting method (M.L.G.-G. *et al.*, unpublished work; http://www.genboree.org). Inspection shows the interleaving of events that both preceded and occurred subsequently to the rat–mouse divergence.

Comparing the three species at 1 Mb resolution, BLASTZ[69], PatternHunter/Grimm-Synteny[70,71], Pash[72], and associated merging algorithms[66,72,73] produce virtually indistinguishable sets of orthologous chromosomal segments. PatternHunter and the GRIMM-Synteny algorithm[73] detect 278 orthologous segments between human and rat, and 280 between human and mouse. The mouse–rat comparison reveals a smaller number of segments (105) of larger average size. The larger number of breaks in orthologous segments between the human to the rodent pair is expected, because of the latter's closer evolutionary relationship.

Understanding the number and timing of rearrangement events that have occurred in each of the three individual lineages (see tree in Fig. 5a) since the common primate–rodent ancestor required a more detailed analysis. We initially focused on the X chromosome, because rearrangements between the X and the autosomes are rare[74] and its history is somewhat easier to trace completely. The X chromosome consists of 16 human–mouse–rat orthologous segments of at least 300 kb in size[73] (Fig. 6a). In the most parsimonious scenario (found with MGR and GRIMM[75]), these were created by 15 inversions in the descent from the primate–rodent ancestor (Fig. 6b). Outgroup data from cat, cow[76] and dog[77] resolved the timing of these rearrangements more precisely. Most of these events occurred in the rodent lineage: five (or four) before the divergence of rat and mouse, five in the rat lineage, and five in the mouse lineage. At most one rearrangement occurred in the human lineage since divergence from the common ancestor with rodents. The timing of this one event was ambiguous, owing to the limited resolution of the outgroup data. Even given this uncertainty, it is clear that the large-scale architecture of the X chromosome in humans is largely unchanged since the primate–rodent ancestor[73], whereas there has been considerable activity in the rodents. The assignment of the accelerated activity to the rodent branch, following the primate–rodent divergence, is consistent with previous studies at significantly lower resolution (these showed complete conservation of marker order between the X chromosomes of human and cat[78], human and dog[77], and human and lemur[79], as well as similar karyotypes of the X chromosomes in human, chimpanzees, gorillas and orangutans[80]).

Large-scale reconstruction of the entire ancestral murid genome suggests that it retained many previously postulated chromosome associations of the placental ancestor[81,82]. The most parsimonious scenario we found requires a total of 353 rearrangements: 247 between the murid ancestor and human, 50 from the murid ancestor to mouse and 56 from the murid ancestor to rat. A recent study[82] implies that most of the 247 rearrangements between the murid ancestor and human occurred on the evolutionary subpath from the squirrel–mouse–rat ancestor to the murid ancestor. Our

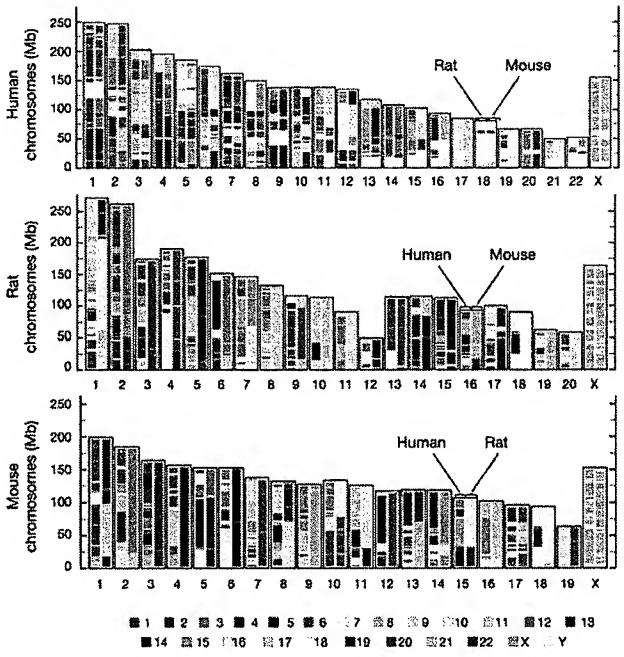
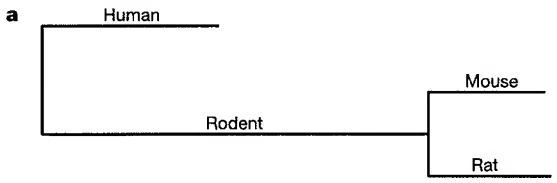


Figure 4 Map of conserved synteny between the human, mouse and rat genomes. For each species, each chromosome (*x* axis) is a two-column boxed pane (p arm at the bottom) coloured according to conserved synteny to chromosomes of the other two species. The same chromosome colour code is used for all species (indicated below). For example, the first 30 Mb of mouse chromosome 15 is shown to be similar to part of human chromosome 5 (by the red in left column) and part of rat chromosome 2 (by the olive in right column). An interactive version is accessible (http://www.genboree.org).



b	Substitutions, insertions and deletions			
	Human	Rodent	Mouse	Rat
Substitutions per site	0.11 ±0.0012	0.24 ±0.0012	0.073 ±0.0014	0.077 ±0.006
Substitutions in neutral sites only	0.13 ±0.011	0.28 ±0.033	0.083 ±0.013	0.091 ±0.011
Insertion events per kb	2.7 ±0.94	4.74 ±1.0	1.54 ±0.84	1.43 ±0.73
Deletion events per kb	5.3 ±0.55	12 ±1.2	3.8 ±0.21	4.5 ±0.13
Inserted bases per kb	6.4 ±2.9	9.4 ±1.6	3.6 ±1.5	3.2 ±1.3
Deleted bases per kb	18 ±2.0	40 ±4.9	11 ±0.55	13 ±0.05

Figure 5 Substitutions and microindels (1–10 bp) in the evolution of the human, mouse and rat genomes. **a**, The lengths of the labelled branches in the tree are proportional to the number of substitutions per site inferred using the REV model[222] from all sites with aligned bases in all three genomes. **b**, The table shows the midpoint and variation in these branch-length estimates when estimated from different sequence alignment programs and different neutral sites, including sites from ancestral repeats[3], fourfold degenerate sites in codons, and rodent-specific sites ('in neutral sites only' row; Supplementary Information). Other rows give midpoints and variation for micro-indels on each branch of the tree in **a**.

analyses confirm that the rate of rearrangements in murid rodents is much higher than in the human lineage[73].

Segmental duplications

Segmental duplications are defined here as regions of the genome that are repeated over at least 5 kb of length and >90% identity. The rat has approximately 2.9% of its bases in these duplicated regions (Fig. 3), whereas the human genome has 5–6%[83]. In contrast to the greater rate of large-scale rearrangement, the mouse genome shows substantially fewer of these events[3], with only 1.0–2.0%[51] of its sequenced bases in duplicated regions. These duplicated structures are particularly challenging to assemble, and we attribute at least some of the mouse–rat differences to the BAC-based approach we used for Rnor3.1, compared with the WGS mouse approach. The vast majority of these sequences (73 of 82 Mb) were regions with <99.5% identity and thus were not simply overlapping sequences that had not been joined by the assembly program Phrap. The 'unplaced' chromosome in Rnor3.1 showed a marked enrichment for blocks of segmental duplication (nearly 44% of the total), which indicates problems with anchoring these elements to the genome.

Intrachromosomal duplications are represented at a three-to-one excess when compared with interchromosomal duplications, and are significantly enriched near the telomeres and in centromeric regions (Fig. 3). The pericentromeric accumulation of segmental duplications in the rat is reminiscent of that observed in human and mouse[83–86], and seems to be a general property of mammalian chromosome architecture.

We observed considerable clustering of duplications[87], including 41 discrete genomic regions larger than 1 Mb in size in which duplications appear to be organized into groups with <100 kb between duplicated segments. For many of these clusters, the underlying sequence alignments showed a wide range in the degree of sequence identity, suggesting that these areas have been subject to duplication events more or less continuously over millions of years. In contrast, an analysis of the evolutionary distance between all duplicated regions showed an unusual bimodal distribution, particularly for intrachromosomal segmental duplications. Two peaks were observed at 0.045 substitutions per site and 0.075 substitutions per site. Given that the rat genome has accumulated 8–10% substitutions (see below) since the speciation from mouse 12–24 Myr ago, this bimodal distribution may correspond to bursts of segmental duplication that occurred approximately 5 and 8 Myr ago, respectively.

The segmental duplications in the rat genome were of considerable interest because they represent an important mechanism for the generation of new genes. We found that 63 NCBI reference sequence[88] (RefSeq; see http://www.ncbi.nih.gov/RefSeq/) genes were located completely or partially within rat duplicated regions, out of a genome total of 4,532 rat RefSeq genes. As discussed below, many of these genes are present in multiple copies and belong to gene familes that have been recently duplicated and contribute to distinctive elements of rat biology.

Gains and losses of DNA

In addition to large rearrangements and segmental duplications, genome architecture is strongly influenced by insertion and deletion events that add and remove DNA over evolutionary time. To characterize the origins and losses of sequence elements in the human, mouse and rat genomes, we categorized all the nucleotides in each of the three genomes, using our alignment data and

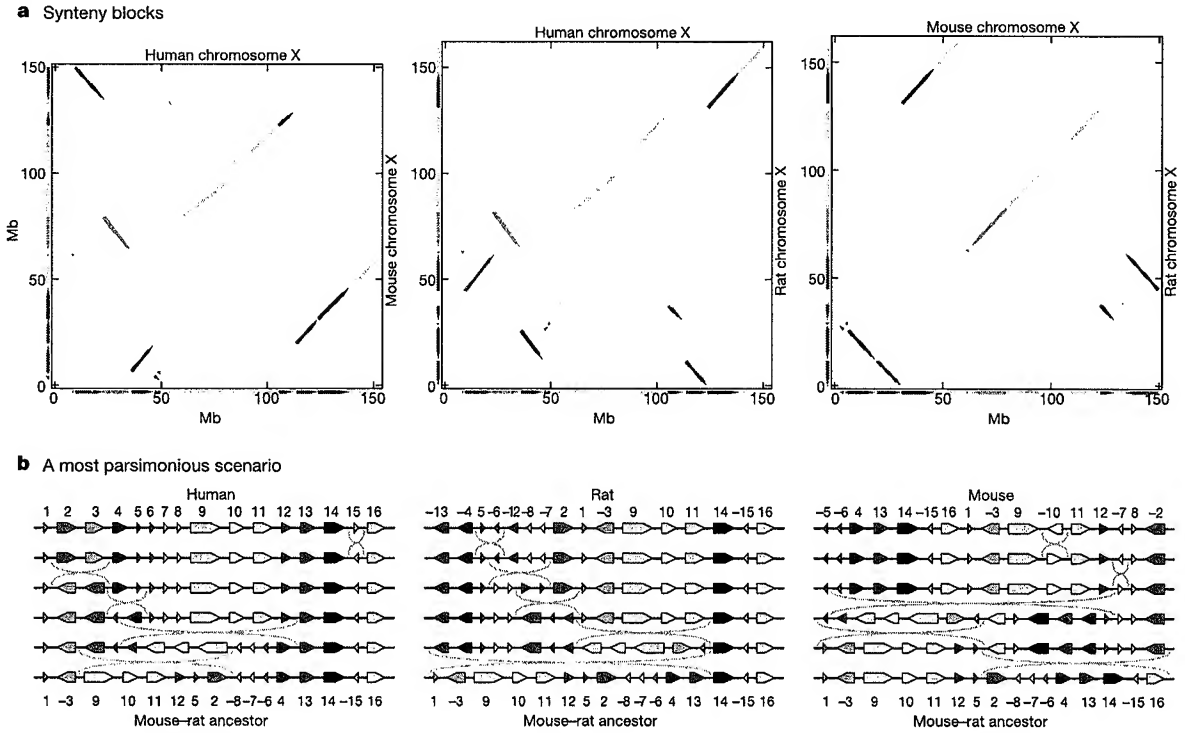


Figure 6 X chromosome in each pair of species. **a**, GRIMM-Synteny[71] computes 16 three-way orthologous segments (≥300 kb) on the X chromosome of human, mouse and rat, shown for each pair of species, using consistent colours. **b**, The arrangement (order and orientation) of the 16 blocks implies that at least 15 rearrangement events occurred during X chromosome evolution of these species. The program MGR (http://www.cs.ucsd.edu/groups/bioinformatics/MGR/) determined that evolutionary scenarios with 15 events are achievable and all have the same median ancestor (located at the last common mouse–rat ancestor). Shown is a possible (not unique) most parsimonious inversion scenario from each species to that ancestor. We note that the last common ancestor of human, mouse and rat should be on the evolutionary path between this median ancestor and human.

RepeatMasker annotations of the insertions of repetitive elements (Fig. 7). The rodent repeat database used by RepeatMasker was greatly expanded by analysing the rat and mouse genomes[89], but it is clear that not all repeats are being recognized, especially the older ones. Thus, these estimates of the amount of rodent repeats represent lower bounds.

About a billion nucleotides (39% of the euchromatic rat genome) align in all three species, constituting an 'ancestral core' that is retained in these genomes. This ancestral core contains 94–95% of the known coding exons and regulatory regions. Comparisons between the human and mouse genomes, using transposon relics retained in both species ('mammalian ancestral repeats') to model neutral evolution, have been used to estimate the fraction of the human genome that is accumulating substitutions more slowly than the neutral rate in both lineages since their divergence, and hence may be under some level of purifying selection[3]. Depending on details of methodology, such estimates have ranged between about 4% and 7%[3,90,91]. The levels of three-way conservation observed here between the human, mouse and rat genomes in the ancestral core lend further support to these earlier estimates, giving values in the range of 5–6% when measured by two quite different methods (see Methods and ref. 92). In this constrained fraction, non-coding regions outnumber coding regions regardless of the strength of constraint[92], an observation that supports recent comparative analyses limited to subsets of the genome[93,94]. The preponderance of non-coding elements in the most constrained fraction of the genome underscores the likelihood that they play critical roles in mammalian biology.

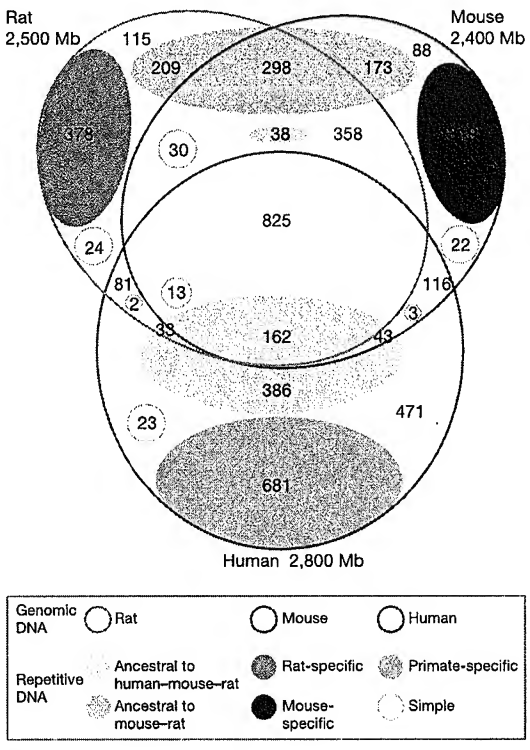


Figure 7 Aligning portions and origins of sequences in rat, mouse and human genomes. Each outlined ellipse is a genome, and the overlapping areas indicate the amount of sequence that aligns in all three species (rat, mouse and human) or in only two species. Non-overlapping regions represent sequence that does not align. Types of repeats classified by ancestry: those that predate the human–rodent divergence (grey), those that arose on the rodent lineage before the rat–mouse divergence (lavender), species-specific (orange for rat, green for mouse, blue for human) and simple (yellow), placed to illustrate the approximate amount of each type in each alignment category. Uncoloured areas are non-repetitive DNA—the bulk is assumed to be ancestral to the human–rodent divergence. Numbers of nucleotides (in Mb) are given for each sector (type of sequence and alignment category). Detailed results are tabulated (Supplementary Table SI-1).

About 700 Mb (28%) of the rat euchromatic genome aligns only with the mouse. At least 40% of this comprises of rodent-specific repeats inserted on the branch from the primate–rodent ancestor to the murid ancestor, and some of the remainder can be recognized as mammalian ancestral repeats whose orthologues were deleted in the human lineage (Fig. 7). Another part is likely to consist of single-copy ancestral DNA deleted in the human lineage but retained in rodents. Although this 700 Mb of rodent-specific DNA is primarily neutral, it may also contain some functional elements lost in the human lineage in addition to sequences representing gains of rodent-specific functions, including some coding exons[95].

The remainder of the euchromatic rat genome (726 Mb, 29%) aligns with neither mouse nor human (Fig. 7). At least half of this (15% of the rat genome) consists of rat-specific repeats, and another large fraction (8% of the rat genome) consists of rodent-specific repeats whose orthologues are deleted in the mouse.

Substitution rates

The alignment data allow relatively precise estimates of the rates of neutral substitutions and microindel events (≤10 bp). Both synonymous fourfold degenerate ('4D') sites in protein-coding regions and sites in mammalian ancestral repeats were used in this analysis, as in previous studies comparing human and mouse[3,96]. We additionally used a class of primarily neutral sites whose identification is made uniquely possible by the addition of the rat genome sequence: namely, the rodent-specific sites discussed above, identified by their failure to align to human sequence.

Our estimates for the neutral substitution level between the two rodents range from 0.15 to 0.20 substitutions per site, while estimates for the entire tree of human, mouse and rat range from 0.52 to 0.65 substitutions per site (Fig. 5). This difference was predictable because of the evolutionary closeness of the two rodents. For all classes of neutral sites analysed, however, the branch connecting the rat to the common rodent ancestor is 5–10% longer than the mouse branch (Fig. 5a). Thus, for as yet unknown reasons, the rat lineage has accumulated substantially more point substitutions than the mouse lineage since their last common ancestor.

We also analysed four-way alignments including sequence from orthologous ancestral repeats in human, mouse and rat, along with the repeat consensus sequences, which approximate the sequence of the progenitor of the corresponding repeat family (Methods). These alignments allow us to distinguish substitutions on the branch from the primate–rodent ancestor to the rodent ancestor from substitutions on the branch descending to human[77]. This revealed an overall speed-up in rodent substitution rates relative to human of about three-to-one, larger than estimated previously[3], but consistent with other more recent studies which also use multiple sequence alignments[77,97,98].

Estimates for rates of microdeletion events are, for all branches, approximately twofold higher than rates of microinsertion (Fig. 5b), suggesting a fundamental difference in the mechanisms that generate these mutations. Furthermore, there are substantial rate differences for each class of event between the various lineages. In particular, the rat lineage has accumulated microdeletions more rapidly than the mouse, while the opposite holds true for microinsertions. As with substitutions, both microinsertion and microdeletion rates are substantially slower in the human lineage. The size distribution of microindels (1–10 bp) on the rat branch was heavily weighted towards the smallest indels: 45% of indels are single bases, 18% are 2 bp, 10% are 3 bp, 8% are 4 bp, and so on, monotonically decreasing. Separate distributions for insertions and for deletions were similar, as were distributions of indel sizes on the mouse branch.

Male mutation bias

As mouse and rat are similar in generation time and number of germline cell divisions[99,100], we investigated a potential sex bias in different types of observed genome changes. We compared substitution and indel rates between the X chromosome and autosomes in ancestral repeat sites (~5 Mb and ~100 Mb in total for X and autosomes, respectively[101]). We discovered that in rodents, small indels (<50 bp) are male-biased, with a male-to-female rate ratio of ~2.3. This is in contrast to a recent study in primates, based on a substantially smaller data set, that indicates no sex bias in small indels[102]. Our male-to-female nucleotide substitution rate ratio in rodents is ~1.9, confirming earlier reports[103,104]. When substitution rates are compared for all sites aligned between mouse and rat (~78 Mb and ~1,691 Mb, respectively), we again observe an approximately twofold excess of small indels and nucleotide substitutions originating in males compared with females[101]. Interestingly, the ratio in the number of cell divisions between the male and female germlines is also about two[99,100], suggesting that these substitutions may arise from mutations that occur primarily during DNA replication.

G+C content and CpG islands

The G+C content of the rat varies significantly across the genome (Fig. 8a), and the distribution more closely resembles that of mouse than human. The variation in G+C content is coupled with differences in the distribution of CpG islands—short regions that are associated with the 5′ ends of genes and gene regulation[2,3,105], and that escape the depletion of CpG dinucleotides that occurs from deamination of methylated cytosine[2,105]. The 2.6 Gb rat genome assembly (including unmapped sequences) contains 15,975 CpG islands in non-repetitive sequences of the genome. This is similar to the 15,500 CpG islands reported in the 2.5 Gb mouse genome[3], but far fewer than the 27,000 reported in the human genome[2,3,105].

A summary of the CpG island distributions by chromosome is given in Fig. 8b. Chromosome X, with a low G+C content of 37.7%, has the fewest islands (362) and the lowest density of islands (2.6 per Mb). Chromosome 12 is at the other end of the range with a G+C content of 43.5% and the highest density of CpG islands (11.5 islands per Mb). This is similar to chromosome 10, with 11.3 islands per Mb. The average density of CpG islands is 5.7 islands per Mb over the whole genome and 5.9 CpG islands per Mb averaged by chromosome, which is similar to the distribution in mouse[3]. Neither rodent genome shows the extreme outliers in CpG island density that are seen for human chromosome 19 (ref. 2). The density of CpG islands in the rat genome correlates positively with the density of predicted genes (R of 0.96) (Fig. 8b).

These data show that the overall changes in CpG island content predate the rat–mouse split and are consistent with the accelerated loss of CpG dinucleotides in rodents compared with humans[105,106]. It remains possible, however, that occurrences such as the greater number of human regions with extremely high G+C content are due to distributional changes mostly in the primate, rather than in the rodent lineage.

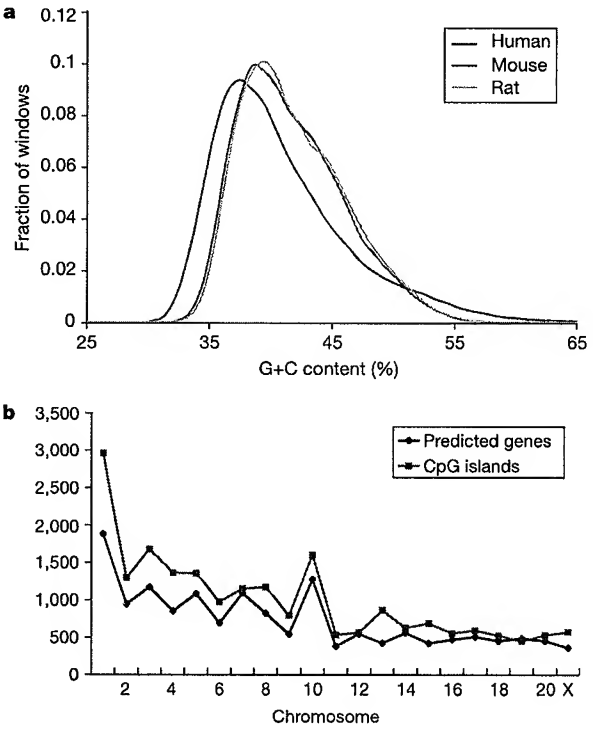


Figure 8 Base composition distribution analysis. **a**, The fraction of 20 kb non-overlapping windows[3] with a given G+C content is shown for human, mouse and rat. **b**, The number of Ensembl-predicted genes per chromosome and the number of CpG islands per chromosome. The density of CpG islands averages 5.9 islands per Mb across chromosomes and 5.7 islands per Mb across the genome. Chromosome 1 has more CpG islands than other chromosomes, yet neither the island density nor ratio to predicted genes exceeds the normal distribution. The number of CpG islands per chromosome and the number of predicted genes are correlated ($R^2 = 0.96$).

Shift in substitution spectra between mouse and rat

The non-repetitive fraction of the rat genome is enriched for G+C content relative to the mouse genome, by ~0.35% over 1.3 billion nucleotides. This is a subtle but substantial difference that may be explained, at least in part, by differences in the spectra of mutation events that have accumulated in the mouse and rat lineages. We analysed all alignment columns in which substitution events can be assigned to either the mouse or the rat lineage, by virtue of a nucleotide match between human and only one rodent[92]; note that this is a small minority of substitutions. Of the ~117 million alignment columns meeting this criteria, ~60 million involve a change in the rat lineage versus ~57 million in the mouse, reflecting the increase in rates of point substitution in the rat lineage (Fig. 5b). While 50% of these changes in rat involve a substitution from an A/T to a G/C, these events constitute only 47% of all mouse changes. The complementary change, G/C to A/T, exhibits relative excess in the mouse versus the rat lineage (38% versus 35%, respectively). No substantial difference between changes that do not alter G+C content is observed. In addition, this bias is not confined to particular transition or transversion events, nor can it be explained simply as a result of divergent substitution rates of CpG dinucleotides (data not shown). Thus, this shift appears to be a general change that results in an increase in G+C content in the rat genome. Biochemical changes in repair or replication enzymes might be responsible, and the observation that recombination rates are slightly higher in rat than in mouse[107] may suggest a role for G+C-biased mismatch repair[108,109]. However, population genetic factors, such as selection, cannot be ruled out.

Evolutionary hotspots

Comparison of the two rodent genomes, using human as outgroup, reveals regions that are conserved yet under different levels of constraint in mouse and rat. These regions may have distinct functional roles and contribute to species-specific differences. Analysis of the MAVID alignments[110] revealed 5,055 regions ≥100 bp, in which there was at least a tenfold difference in the estimated number of substitutions per site on the mouse and rat branches. To avoid alignment problems and fast-evolving regions, the analysis was restricted to regions where the human branch had <0.25 substitutions per site[111]. These regions are enriched twofold in transcribed regions: 39% of mouse hotspots were found in the

18% of the mouse genome covered by RefSeq genes; and 17% of the rat hotspots were found in the 8% of the rat genome covered by RefSeq genes. Similar numbers are observed when examining coding exon and EST regions (not shown). Half of all hotspots in the mouse genome lie totally in non-coding regions. Many hotspots are several hundred bases long, with average length 190 ± 86 bp. Future work aimed at identifying the genomic differences that contribute to phenotypic evolution may benefit from analyses such as these, which will become more powerful as the repertoire of mammalian genome sequences expands.

Covariation of evolutionary and genomic features

To illustrate the genomic and evolutionary landscape of a single rat chromosome in depth, we characterized features for rat chromosome 10 at 1 Mb resolution (Fig. 9). This high-resolution analysis uncovered strong correlations between certain microevolutionary features[89,92,98]. Particularly strongly correlated are the local rates of microdeletion ($R^2 = 0.71$; Fig. 9a), microinsertion ($R^2 = 0.56$; Fig. 9a), and point substitution ($R^2 = 0.86$; Fig. 9b) between the two independent lineages of mouse and rat. In addition, microinsertion rates are correlated with microdeletion rates ($R^2 = 0.55$; Fig. 9a). These strong correlations are also observed in an independent genome-wide analysis, both on the original data and after factoring out the effects of G+C content (not shown, see Supplementary Information).

Perhaps surprisingly, substantially less correlation is seen between microindel and point substitution rates (compare Fig. 9a and b). The amount of correlation varies among chromosomes (not shown), but is generally weaker than the relationships mentioned above. Further studies will be required to determine whether local evolutionary pressures, which must have remained stable since the separation of the mouse and rat lineages, differentially drive microindel and point substitution rates.

We also find that the local point substitution rate in sites common to human, mouse and rat strongly correlates with that in rodent-specific sites ($R^2 = 0.57$; Fig. 9b, blue line versus red/green). These two classes of sites, while interdigitated at the level of tens to thousands of bases, constitute sites that are otherwise evolutionarily independent. This result confirms that local rate variation is not solely determined by stochastic effects and extends, at high resolution, the previously documented regional correlation in rate between 4D sites and ancestral repeat sites[3,96].

Evolution of genes

A substantial motivation for sequencing the rat genome was to study protein-coding genes. Besides being the first step in accurately defining the rat proteome, this fundamental data set yields insights into differences between the rat and other mammalian species with a complete genome sequence. Estimation of the rat gene content is possible because of relatively mature gene-prediction programs and rodent transcript data. Mouse and human genome sequences also allow characterization of mutational events in proteins such as amino acid repeats and codon insertions and deletions. The quality of the rat sequence also allows us to distinguish between functional genes and pseudogenes.

We estimate (on the basis of a subset) that 90% of rat genes

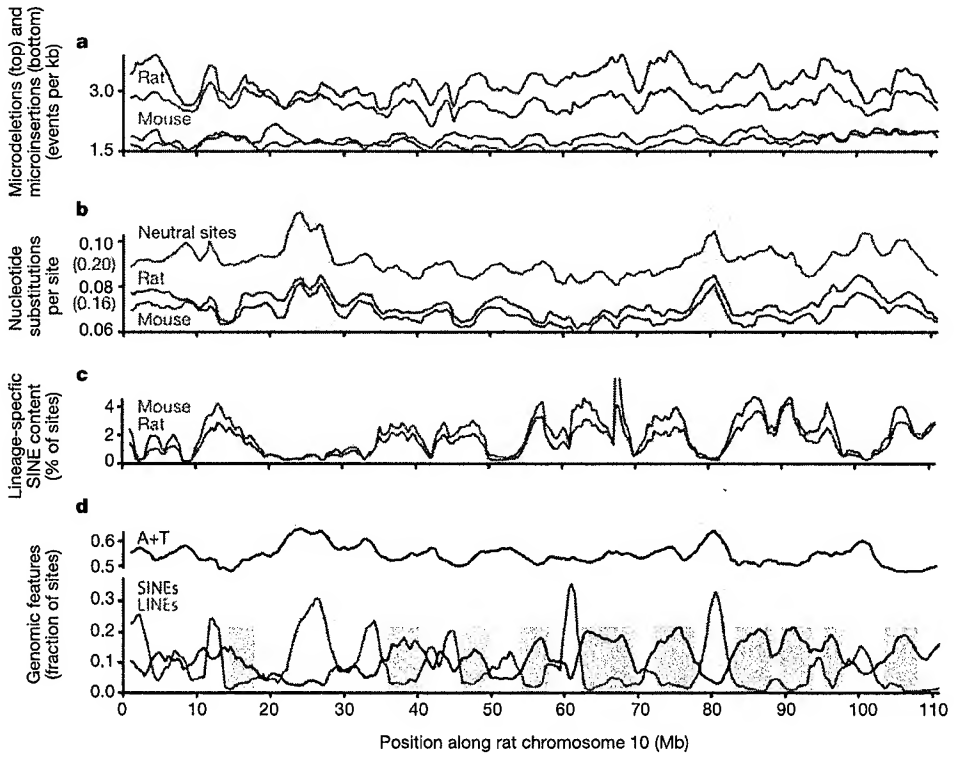


Figure 9 Variability of several evolutionary and genomic features along rat chromosome 10. **a**, Rates of microdeletion and microinsertion events (less than 11 bp) in the mouse and rat lineages since their last common ancestor, revealing regional correlations. **b**, Rates of point substitution in the mouse and rat lineages. Red and green lines represent rates of substitution within each lineage estimated from sites common to human, mouse and rat. Blue represents the neutral distance separating the rodents, as estimated from rodent-specific sites. Note the regional correlation among all three plots, despite being estimated in different lineages (mouse and rat) and from different sites (mammalian versus rodent-specific). **c**, Density of SINEs inserted independently into the rat or mouse genomes after their last common ancestor. **d**, A+T content of the rat, and density in the rat genome of LINEs and SINEs that originated since the last common ancestor of human, mouse and rat. Pink boxes highlight regions of the chromosome in which substitution rates, A+T content and LINE density are correlated. Blue boxes highlight regions in which SINE density is high but LINE density is low.

possess strict orthologues in both mouse and human genomes. Our studies also identified genes arising from recent duplication events occurring only in rat, and not in mouse or human. These genes contribute characteristic features of rat-specific biology, including aspects of reproduction, immunity and toxin metabolism. By contrast, almost all human 'disease genes' have rat orthologues. This emphasizes the importance of the rat as a model organism in experimental science.

Construction of gene set and determination of orthology

The Ensembl gene prediction pipeline[112] predicted 20,973 genes with 28,516 transcripts and 205,623 exons (Methods). These genes contain an average of 9.7 exons, with a median exon number of 6.0. At least 20% of the genes are alternatively spliced, with an average of 1.3 transcripts predicted per gene. Of the 17% single exon transcripts, 1,355 contain frameshifts relative to the predicted protein and 1,176 are probably processed pseudogenes. Of the 28,516 transcripts, 48% have both 5′ and 3′ untranslated regions (UTRs) predicted and 60% have at least one UTR predicted.

These gene predictions considered homology to other sequences, including 26,949 rodent proteins, 4,861 non-rodent, vertebrate proteins, 7,121 rat complementary DNAs from RefSeq and EMBL, and 31,545 mouse cDNAs from Riken, RefSeq and EMBL. The majority (61%) of transcripts are supported by rodent transcript evidence. When combined with additional private EST data, the fraction of genes supported by transcript evidence could be increased to 72%[113].

A number of other *ab initio* (GENSCAN[114], GENEID[115]), similarity-based (FGENESH++; ref. 116) and comparative (SGP[117], SLAM[118], TWINSCAN1[119–121]) gene-prediction programs were used to analyse the rat genome. The number of genes predicted by these programs ranged from 24,500 to 47,000, suggesting coding densities ranging from 1.2% to 2.2%. The coding fraction of RefSeq genes covered by these predictions ranged from 82% to 98%. Such comparative *ab initio* programs using the rat genome were successfully used to identify and experimentally verify genes missed by other methods in rat[121] and human[122]. The predictions of these programs can be accessed through the UCSC genome browser and Ensembl websites.

RefSeq genes (20,091 human, 11,342 mouse and 4,488 rat) mapped onto genome assemblies with BLAT[123] and the UCSC browser revealed that the number of coding exons per gene and average exon length were similar in the three species. Differences were observed in intron length, with an average of 5,338 bp in human, 4,212 bp in mouse and 5,002 bp in rat. These differences were also found in a smaller collection of 6,352 confidently mapped orthologous intron triads (see 'Conservation of intronic splice signals' section below): average intron lengths in this collection were 4,240 bp in human, 3,565 bp in mouse and 3,638 bp in rat.

Properties of orthologous genes

Orthology relationships were predicted on the basis of BLASTp reciprocal best-hits between proteins of genome pairs (human–rat, rat–mouse and mouse–human)[3] (Supplementary Information). Using these methods and the ENSEMBL prediction sets, 12,440 rat genes showed clear, unambiguous 1:1 correspondence with a gene in the mouse genome. This is an underestimate, because random sampling of different classes of rat genes with less stringent criteria for comparison to mouse always identified additional gene pairs. Errors arose from pseudogene misclassification, sequence loss, duplication or fragmentation in assemblies; and missing or inappropriate gene predictions, including coding-gene predictions from non-coding RNAs. Taking these errors into account, we estimate the true proportion of 1:1 orthologues in rat and mouse genomes to lie between 86 and 94% (Methods). The remaining genes were associated with lineage-specific gene family expansions or contractions. These overall observations are consistent with a careful analysis of rat proteases showing that 93% of these genes have 1:1 orthologues in mouse[124,125].

Surprisingly, a similar proportion (89 to 90%) of rat genes possessed a single orthologue in the human genome. Because human represents an outgroup to the two rodents, it was expected that mouse and rat would share a higher fraction of orthologues. A close inspection of gene relationships indicates that these findings may suffer from incompleteness of rodent genome sequences, together with problems of misassembly and gene prediction within clusters of gene paralogues.

Further analysis of orthologous pairs considered the occurrence of nucleotide changes within protein-coding regions that reflected synonymous or non-synonymous substitutions. The majority of these studies measured evolutionary rates by determination of K_A (number of non-synonymous substitutions per non-synonymous site) and K_S (number of synonymous substitutions per synonymous site). K_A/K_S ratios of less than 0.25 indicate purifying selection, values of 1 suggest neutral evolution, and values greater than 1 indicate positive selection[126].

Evolutionary rates were first calculated from a reduced set of orthologue pairs that are embedded in orthologous genomic segments and are related by conservative values of K_S (Table 3) (Methods). A slight increase in median K_S values for rat–human as compared with mouse–human, was found, indicating that the rat lineage has more neutral substitutions in gene coding regions than the mouse lineage. Sequence conservation values were similar to those previously found using smaller data sets[127,128], and the overall trend is consistent with results of other evolutionary rate analyses discussed above (Fig. 5).

Next, we investigated examples of rat genes shared with mouse, but with no counterparts in human. Such genes might be rapidly evolving so that homologues are not discernible in human, or they might have arisen from non-coding DNA, or their orthologues in the human lineage might have formed pseudogenes. Thirty-one Ensembl rat genes were collected that have no non-rodent homologues in current databases (Methods). These are twofold over-represented among genes in paralogous gene clusters, and threefold over-represented among genes whose proteins are likely to be secreted. This is consistent with observations[3] that clusters of paralogous genes, and secreted proteins, evolve relatively rapidly. Detailed examination of the 31 genes using PSI-BLAST determined that ten genes cannot be assigned homology relationships to experimentally described mammalian genes. These ten rodent-

Table 3 **One-to-one orthologous genes in human, mouse and rat genomes**

	Human–mouse	Human–rat	Mouse–rat
1:1 orthologue relationships	11,084	10,066	11,503
Median K_S values*	0.56 (0.39–0.80)	0.57 (0.40–0.82)	0.19 (0.13–0.26)
Median K_A/K_S values*	0.10 (0.03–0.24)	0.09 (0.03–0.21)	0.11 (0.03–0.28)
Median % amino acid identity*	88.0% (74.4–96.3%)	88.3% (75.9–96.4%)	95.0%† (88.0–98.7%)
Median % nucleotide identity*	85.1% (77.4–90.0%)	85.1% (77.8–89.9%)	93.4% (89.2–95.7%)

Data obtained from Ensembl, *Homo sapiens* version 11.31 (24,841 genes), *Mus musculus* version 10.3 (22,345 genes), *Rattus norvegicus* version 11.2 (21,022 genes).
*Numbers in parentheses represent the 16th and 83rd percentiles.
†This value is consistent with previous findings (93.9% in ref. 130).

specific genes may have evolved particularly rapidly, or have non-coding DNA homologues, or be erroneous predictions.

The paucity of rodent-specific genes indicates that *de novo* invention of complete genes in rodents is rare. This is not unexpected, because the majority of eukaryotic protein-coding genes are modular structures containing coding and non-coding exons, splicing signals and regulatory sequences, and the chances of independent evolution and successful assembly of these elements into a functional gene are small, given the relatively short evolutionary time available since the mouse–rat split. However, individual rodent-specific exons may arise more frequently, particularly if the exon is alternatively spliced[129]. Applying a K_A/K_S ratio test[130,131] to sequences that align only between rat and mouse, we identified 2,302 potential novel rodent-specific exons, with EST support, in BLASTZ alignments of rat and mouse sequences. None of these individual exons matched human transcripts, but approximately half (1,116) appear to be present in alternative splice forms found in rodents. We speculate that these exons contain the few successful lineage-specific survivors of the constant process of gene evolution, by birth and death of individual exons.

Indels and repeats in protein-coding sequences

In contrast to small indels occurring in the bulk of the genome (above), indels within protein-coding regions are probably lethal, or deleterious and so are rapidly removed from the population by purifying selection. Indel rates within rat coding sequences were 50-fold lower than in bulk genomic DNA[132]. The whole genome excess of deletions compared with insertions (Fig. 5b) was also evident in coding sequences. The magnitude was less, with a genome-wide deletion-to-insertion ratio of 3.1:1 reducing to 1.7:1 in the rat. In mouse this value reduced from 2.5:1 to 1.1:1 (ref. 132). These data suggest that deletions are ~16% more likely than insertions to be removed from coding sequences by selection.

Owing to the triplet nature of the genetic code, indels of multiples of three nucleotides in length (3_n indels) are less likely to be deleterious. Direct comparison of 3_n indel rates between bulk DNA (0.77 indels per kb for mouse, 0.83 indels per kb for rat) and coding sequence (0.087 indels per kb for mouse and 0.084 indel per kb for rat) showed that 3_n indels were ninefold under-represented in coding sequences. At least 44% of indels were duplicative insertion or deletion of a tandemly duplicated sequence, collectively termed sequence slippage[132]. Sequence slippage contributed approximately equally to observed insertions and deletions. The overall excess of deletions could be attributed specifically to an excess of non-slippage deletion over non-slippage insertion in both mouse and rat lineages[132]. Of the slippage indels, 13% were in the context of trinucleotide repeats ($n > 2$, excluding the inserted or deleted sequence) which are known to be particularly prone to sequence slippage and encode homopolymeric amino acid tracts[133,134].

To gain better understanding of dynamic changes in the length of homopolymeric amino acid tracts on gene evolution and disease susceptibility, we searched for other characteristics of amino acid repeat variation by analysing all size-five or longer amino acid repeats in a data set of 7,039 rat, mouse and human orthologous protein sequences[135]. Most species-specific amino acid repeats (80–90%) were found in indel regions, and regions encoding species-specific repeats were more likely to contain tandem trinucleotide repeats than those encoding conserved repeats. This was consistent with the involvement of slippage in the generation of novel repeats in proteins and extended previous observations for glutamine repeats in a more limited human–mouse data set[136].

The percentage of proteins containing amino acid repeats was 13.7% in rat, 14.9% in mouse and 17.6% in human[135]. The most frequently occurring tandem amino acid repeats were glutamic acid, proline, alanine, leucine, serine, glycine, glutamine and lysine. Using the same threshold size cut-off, tandem trinucleotide repeats were significantly more abundant in human than in rodent coding sequences, in striking contrast to the frequencies observed in bulk genomic sequences (29 trinucleotide repeats per Mb in rat, 32 repeats per Mb in mouse and 13 repeats per Mb in human, see discussion of the general simple repeat structure below). The conservation of human repeats was higher in mouse (52%) than in rat (46.5%), suggesting a higher rate of repeat loss in the rat lineage than the mouse lineage.

Functional consequences of these in-frame changes in rat, mouse and human were investigated[132] through clustering of proteins based on annotation of function and cellular localization[112], and mapping indels onto protein structural and sequence features. The rate that indels accumulated in secreted (3.9×10^{-4} indels per amino acid) and nuclear (4.0×10^{-4}) proteins is approximately twice that of cytoplasmic (2.4×10^{-4}) and mitochondrial (1.4×10^{-4}) proteins. Likewise, ligand-binding proteins acquire indels (3.1×10^{-4}) at a higher rate than enzymes (2.1×10^{-4})[132]. These trends exactly mirror those observed for amino acid substitution rates[3], suggesting tight coupling of selective constraints between indels and substitutions. Transcription regulators showed the highest rate of indels (4.3×10^{-4}), a finding that may relate to the over-representation of homopolymorphic amino acid tracts in these proteins[135].

Known protein domains exhibited 3.3-fold fewer indels than expected by chance, again paralleling nucleotide substitution rate differences between domains and non-domain sequences[3]. Of the protein-sequence and structural categories considered (transmembrane, protein domain, signal peptide, coiled coil and low complexity), the transmembrane regions were the most refractory to accumulating indels, exhibiting a sixfold reduction compared with that expected by chance. Low-complexity regions were 3.1-fold enriched, reflecting their relatively unstructured nature and enrichment in indel-prone trinucleotide repeats. Mapping of indels onto groups of known structures revealed that indels are 21% more likely to be tolerated in loop regions than the structural core of the protein[132].

We observed that indel frequency and amino acid repeat occurrence both correlated positively with the G + C coding sequence content of the local sequence environment[132,135]. This may be explained in part by the correlation of polymerase slippage-prone trinucleotide repeat sequences and G + C content[135]. There is also a positive correlation between CpG dinucleotide frequency and coding sequence insertions, but not deletions. This effect diminishes rapidly with increasing distance from the site of the insertion[132].

Transcription-associated substitution strand asymmetry

A recent study reported a significant strand asymmetry for neutral substitutions in transcribed regions[133]. Within introns of nine genes, the higher rate of A→G substitutions over that of T→C substitutions, together with a smaller excess of G→A over C→T substitutions, leads to an excess of G+T over C+A on the coding strand (also verified on human chromosome 22). The authors[133] hypothesized that the asymmetries are a byproduct of transcription-

Table 4 **Strand asymmetry of substitutions in introns of rat genes**

Base frequencies on coding strand*	Rat genome	
(G+T)/(C+A)	1.060	
Ratio of purine transitions to pyrimidine transitions†	Rat–mouse	Rat–human
Rate(A↔G)/Rate(C↔T)	1.036	1.036
Rate of transitions‡	Rat	Mouse
Rate(A→G)/Rate(T→C)	1.058	1.091
Rate(G→A)/Rate(C→T)	1.017	1.00

*Computed from the rat genome.
†Computed from pairwise alignments.
‡Computed from three-way alignments.

coupled repair in germline cells. Examining the three-way alignments of rat, mouse and human, we verified that the strand asymmetries for neutral substitutions exist in introns across the genome (Table 4).

Under the assumption of independence of sequence positions, large sample normal approximations to the binomial distribution allow us to test whether the fraction of G+T exceeds 0.5, and whether the rate at the numerator exceeds the rate at the denominator for each of the ratios in Table 4. With the large amount of data provided by pooling introns genome-wide, the tests are all highly significant (P values $< 10^{-4}$), except for the rate of G→A in mouse, which does not significantly exceed that of C→T (P value = 0.6369). These asymmetries are also seen if the study is limited to ancestral repeat sites, excludes ancestral repeat sites, excludes CpG dinucleotides, is limited to positions flanked by sites that are identical in the aligned sequences (in the case of observations 2 and 3 in Table 4), or considers introns of RefSeq genes for human or mouse. Thus it appears that strand asymmetry of substitution events within transcribed regions of the genome is a robust genome-wide phenomenon.

Conservation of intronic splice signals

Using 6,352 human–mouse–rat orthologous introns from 976 genes (Methods), we examined the dynamics of evolution of consensus splice signals in mammalian genes. We found that intron class[137] is extremely well conserved: we did not observe any U2 to U12 intron conversion, or vice versa, nor within U12 introns did we find any switching between the major AT–AC and GT–AG subtypes, although such events are documented at larger evolutionary distances[137]. In contrast, conversions between canonical GT–AG and non-canonical GC–AG subtypes of U2 introns are not uncommon. Only ~70% of GC–AG introns are conserved between human and mouse/rat, and only 90% are conserved between mouse and rat. Using human as the outgroup, we detected nine GT to GC conversions after divergence of mouse and rat (from 6,282 introns that were likely to have been GT–AG before human and rodents split), and two GC to GT conversions (from 34 GC–AG introns that probably predated the human and rodent split). These results give some indication of the degree to which mutation from T to C is tolerated in donor sites. The GC donor site appears to be better tolerated in introns with very strong donor sites, because in these introns the proportion of GC donor sites is ~11%, much higher than the 0.7% overall frequency of GC donor sites in U2 introns. Although we found a variety of other non-canonical configurations in U2 introns, very few are conserved, which suggests that most correspond to transient, evolutionarily unstable states, pseudogenes, or mis-annotations.

Gene duplications

Duplication of genomic segments represents a frequent and robust mechanism for generating new genes[138]. Because there were no compelling data showing rat-specific genes arising directly from non-coding sequences, we examined gene duplications to measure their potential contribution to rat-specific biology. A previous study showed that gene clusters in mouse without counterparts in human are subject to rapid, adaptive evolution[3,139]. We used two methods to identify recent gene duplications: methods that directly identified paralogous clusters, and methods that analysed genomic segmental duplications (see above).

Using the first approach, we found 784 rat paralogue clusters containing 3,089 genes (Methods). This was lower than in mouse (910 clusters/3,784 genes), but the difference probably reflects the larger number of gene predictions from the mouse assembly.

To investigate the timing of expansion of these individual families, we measured rates of local gene duplication and retention within clusters. BLAST is not suited to this[140,141] and so we instead calculated the number of synonymous substitutions per synonymous site (K_S) between all pairs of homologous genes; constructed K_S-derived phylogenetic trees; and predicted orthology or paralogy gene duplication events automatically from their topologies (Supplementary Information). The results showed that the neutral substitution rate varies among orthologues by approximately twofold (Fig. 10). This is similar to chromosomal variation shown previously by a study of mouse and human ancestral repeats[3]. Rates of change among ancestral gene duplications (those that predate the mouse–rat split) were relatively constant. Mouse-specific and rat-specific duplications occurred at similar rates, except for those with $K_S < 0.04$, which are reduced in mouse-specific duplications (Fig. 10). More data are required to determine whether this reduction is a biological effect, as it might be accounted for by different protocols for assembling mouse and rat genomes, which differentially collapse areas of nearly identical sequence.

The rat paralogue pairs that probably arose after the rat–mouse split (12–24 Myr ago) have K_S values of ≤0.2 (Table 3). We found 649 $K_S < 0.2$ gene duplication events in rat, a lower number than is found in mouse (755). For both rodents, this represents a likelihood of a gene duplicating of between 1.3×10^{-3} and 2.6×10^{-3} every Myr. These are necessarily estimates, because gene deletions, conversions and pseudogene formation are not considered. Interestingly, the data are consistent with a previous estimate for *Drosophila* genes, but are an order of magnitude lower than an estimate for *Caenorhabditis elegans* genes[140].

A subset of clusters have at least three gene duplications with $K_S < 0.2$ (Table 5). These are expected to be enriched in genes whose duplications persist as a consequence of positive selection. The group is dominated by genes involved in adaptive immune response and chemosensation[87]. Inspection of the K_S-derived trees allowed us to infer the gene numbers in these clusters for the common ancestor of rat and mouse (that is, at $K_S = 0.2$), assuming no gene deletions or pseudogene generation (Table 5). Immunoglobulin, T-cell receptor α-chain, and α_{2u}-globulin genes appear to be duplicating at the fastest rates in the rat genome (Table 5). Since divergence with mouse, these rat clusters have increased gene content several-fold. This recapitulates previous observations that rapidly evolving and duplicating genes are over-represented in olfaction and odorant detection, antigen recognition and reproduction[142].

An examination of duplicated genomic segments showed this enrichment for most of the same genes and also elements involved in foreign compound detoxification (cytochrome P450 and carboxylesterase genes)[87]. Together, these are exciting findings because each of these categories can easily be associated with a

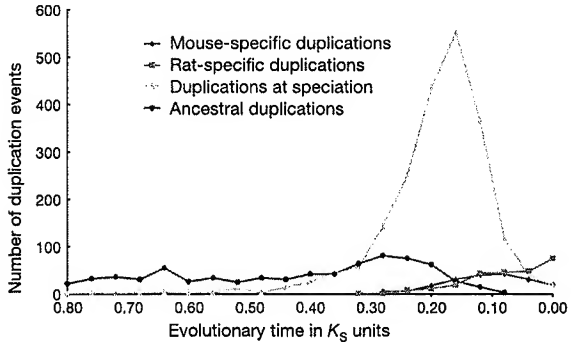


Figure 10 Variation in the frequency of gene duplications during the evolutionary histories of the rat and mouse. The sequence of gene duplication events was inferred from phylogenetic trees determined from pairwise estimates of genetic divergence under neutral selection (K_S, Methods). The median K_S value for mouse:rat 1:1 orthologues is 0.19. This value corresponds to the divergence time of mouse and rat lineages.

familiar feature of rat-specific biology, and further investigation could explain some differences between rats and their evolutionary neighbours.

Conservation of gene regulatory regions

As the third mammal to be fully sequenced, the rat can add significantly to the utility of nucleotide alignments for identifying conserved non-coding sequences[143–147]. This power increases roughly as a function of the total amount of neutral substitution represented in the alignment[97,98], and rat adds about 15% to the human–mouse comparison (Fig. 5). Many conserved mammalian non-coding sequences are expected to have regulatory function, and can be predicted using further analyses based upon these alignments[93,148–150].

We applied such methods for detecting significantly conserved elements[97,151] and scoring regulatory potential[148,152] to the genome-wide human–mouse–rat alignments. Typical results show strong conservation for a coding exon, as well as for several non-coding regions (Fig. 11). For example, the intronic region in Fig. 11 contains 504 bp that are highly conserved in human, mouse and rat. The last 100 bp of this alignment block are identical in all three species. Peaks in regulatory potential score are correlated with conservation score, and in the highly conserved intronic segment, they are higher for the three-way regulatory potential score than for the two-way scores using human and just one rodent[152]. These data are illustrative, but form the foundation of ongoing efforts to identify genome sequences involved in gene regulation.

Requiring conservation among mammalian genomes greatly increases the specificity of predictions of transcription factor binding sites. Transcription factor databases such as TRANSFAC[153] contain known transcription factor binding sites and some knowledge of their distribution, but simply searching a sequence with these motifs provides little discriminatory power. For example, all of the 85 known regulatory elements[148] and 151 functional promoters[154] have TRANSFAC matches, but so do 99% of the 2,049,195 mammalian ancestral repeats, most representing false-positive predictions. The introduction of conservation as a criterion for regulatory element identification greatly increases specificity, with only a modest cost in sensitivity. If we insist that the TRANSFAC matches be present and orthologously aligned in all three species—human, mouse and rat—then only 268 matches are recorded in ancestral repeats (0.01%), while 63 (74%) of the above matches in known regulatory elements and 121 (80%) in functional promoters are retained. Overall, using a set of 164 weight matrices for 109 transcription factors extracted from TRANSFAC[153], we find 186,792,933 matches in the April 2003 reference human genome sequence, but this was reduced to only 4,188,229 by demanding conservation in the human–mouse–rat three-way alignments. This is a 44-fold increase in specificity.

We examined one region in more detail: a complex *cis*-regulatory region consisting of a 4,000 bp segment containing two regulatory modules, hypersensitive sites 2 and 3 from the locus control region of the HBB complex[155–157]. Considerable experimental work has identified six functional binding sites for the transcription factor GATA-1 in this segment. Requiring that matches to GATA-1 binding sites be conserved in all three species and occur within regions of strong regulatory potential is sufficient to find these six functional binding sites, and only these six, in the 4,000 bp segment. Thus, in this example we observed complete sensitivity and specificity by requiring this level of conservation.

Pseudogenes and gene loss

To complement the identification and analysis of protein-coding regions, we sought to examine rat pseudogenes. Using a previously described method[158,159], we found 18,755 pseudogenes in intergenic regions. Pseudogenes are normally not subjected to selective con-

Table 5 **Recent gene duplications (K_S < 0.2) in the rat lineage**

Cluster ID	Recent duplication events	Numbers of genes involved	Extant cluster size	Ancestral cluster size	Chromosome	Annotation	Process
249	38	53	60	22	4	Immunoglobulin κ-chain V	Immunity
640	38	47	53	15	15	TCR α-chain V	Immunity
346	25	35	44	15	6	Immunoglobulin heavy chain V	Immunity
190	22	42	168	146	3	Olfactory receptor	Chemosensation
578	16	28	59	43	13	Olfactory receptor	Chemosensation
400	15	26	82	67	8	Olfactory receptor	Chemosensation
743	15	21	37	22	20	Olfactory receptor	Chemosensation
72	12	22	102	90	1	Olfactory receptor	Chemosensation
500	12	18	32	20	10	Olfactory receptor	Chemosensation
51	6	7	16	10	1	Glandular kallikrein	Reproduction?
256	6	8	10	4	4	Vomeronasal receptor V1R	Chemosensation
488	6	10	11	5	10	Olfactory receptor	Chemosensation
644	6	10	14	8	15	Granzyme serine protease	Immunity
4	5	6	9	4	1	Trace amine receptor, GPCR	Neuropeptide receptors?
248	5	9	15	10	4	Vomeronasal receptor V1R	Chemosensation
393	5	10	31	26	8	Olfactory receptor	Chemosensation
522	5	8	19	14	10	Keratin-associated protein	Epithelial cell function
550	5	8	17	12	11	Olfactory receptor	Chemosensation
635	5	9	20	15	15	Olfactory receptor	Chemosensation
79	4	8	38	34	1	Olfactory receptor	Chemosensation
88	4	6	11	7	1	Olfactory receptor	Chemosensation
109	4	7	43	39	1	Olfactory receptor	Chemosensation
294	4	5	5	1	5	α_{2u}-globulin	Chemosensation
310	4	5	11	7	5	Olfactory receptor	Chemosensation
353	4	7	13	9	7	Olfactory receptor	Chemosensation
399	4	5	6	2	8	Ly6-like urinary protein	Chemosensation?
638	4	6	6	2	15	RNase A	Immunity
690	4	6	21	17	17	Prolactin paralogue	Reproduction
239	3	6	6	3	4	Prolactin-induced protein	Reproduction
253	3	4	5	2	4	Camello-like *N*-acetyltransferase	Developmental regulator
274	3	6	20	17	4	Ly-49 lectin natural killer cell protein	Immunity
297	3	4	5	2	5	Interferon-α	Immunity
523	3	4	6	3	10	Keratin-associated protein	Epithelial cell function
746	3	5	6	3	20	MHC class 1b (M10)	Chemosensation

Duplications involving retroviral genes, fragmented genes with internal repeats, and likely pseudogene clusters were removed from this list. Only gene clusters exhibiting at least three duplications are shown.

straint and therefore accumulate sequence modifications neutrally. Indeed, nearly all of our identified pseudogenes (97 ± 3%) evolved under neutrality according to a K_A/K_S test, and therefore are consistent with being pseudogenic.

We classified these pseudogenes according to whether they arose from retrotransposition, in which case they integrated into the genome randomly, or whether they arose from tandem duplication and neutral sequence substitution. Using human–rat synteny, we found that 80% of pseudogenes exhibited no significant similarity to the corresponding human orthologous region, and therefore were considered retrotransposed, processed pseudogenes. The total pseudogene count, and processed pseudogene proportion, are consistent with those found for human[158,159]. These numbers are greater than those previously reported for mouse[3,4]. However, reanalysis using the method employed here detects a similar pseudogene number (20,000) to that found for human and rat. This suggests that the rate of pseudogene creation is similar among these mammals.

As with the human genome[159,160], the largest group of rat pseudogenes (totalling 2,188), according to InterPro[161], consists of ribosomal protein genes. Other large rat pseudogene families arose from olfactory receptors (552, see below), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (251), protein kinases (177), and RNA binding RNP-1 proteins (174). Pseudogenes homologous to a meiotic spindle-associated protein—spindlin[162]—are particularly numerous in rat (at least 53 copies) compared with mouse (approximately three copies). This suggests that spindlin pseudogenes may have distributed rapidly by a recently active transposable element.

We investigated the much-studied metabolic enzyme GAPDH[3,163], and observed that: (1) the GAPDS gene arose from a duplication of the GAPDH gene; (2) biogenesis of the GAPDH pseudogenes has been occurring steadily over time both before and after rodent–human and mouse–rat divergence; and (3) the GAPDS gene has undergone little retrotransposition in all three genomes compared with its relative, the GAPDH gene (consistent with respective gene-expression levels in the germ line).

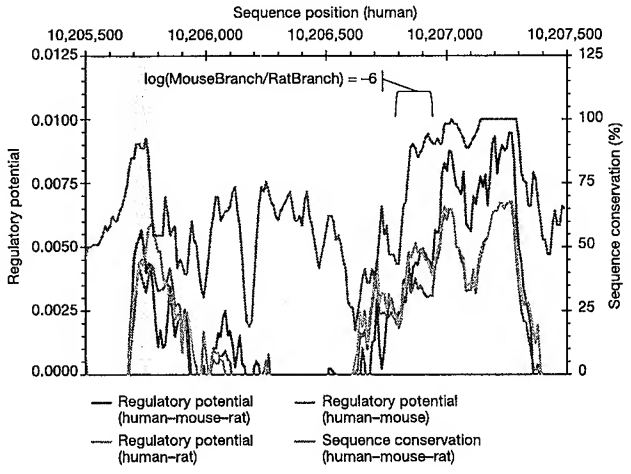


Figure 11 Close-up of PEX14 (peroxisomal membrane protein) locus on human chromosome 1 (with homologous mouse chromosome 4 and rat chromosome 5). Conservation score computed on three-way human–mouse–rat alignments (parsimony *P* values[151]) presents a clear coding exon peak (grey bar) and very high values in a 504 bp non-coding, intronic segment (right; last 100 bp of alignment are identical in all three organisms). The latter segment showed a striking difference between the inferred mouse and rat branch lengths[110,111,222]: the grey bracket corresponds to a phylogenetic tree where the logarithm of mouse to rat branch-length ratio is −6. Regulatory potential scores[148,152] that discriminate between conserved regulatory elements and neutrally evolving DNA are calculated from three-way (human–mouse–rat) and two-way (human–rodent) alignments. Here the three-way regulatory potential scores are enhanced over the two-way scores.

In situ loss of rat genes

As an organism evolves, its need for certain genes may be reduced, or lost, owing to changes in its ecological niche. Loss of selective constraints leads to accumulation of nonsense and/or frameshift mutations without retrotransposition or duplication. These non-processed pseudogenes are interesting because they link environmental changes to genomic mutation events. However, predicted pseudogenes with disrupted reading frames might also be indicative of errors in genome sequence or assembly. By constraining the search to orthologous genomic regions, we identified 14 rat putative non-processed pseudogenes (Table 6) with apparently functional, single human and mouse orthologues. Half of these contain one in-frame stop or frameshift, whereas the remainder contain more. We expect this number of identified pseudogenic orthologues to be conservative because the methods employed required high fidelity of both gene prediction and orthologue identification in all three species (Methods).

Nevertheless, as only 14 recently evolved pseudogene candidates were identified, this indicates that the genome sequence and assembly (Rnor3.1) is of high quality. The improved quality of the most recent assembly is underscored by 11 additional candidate pseudogenes, predicted from rat assembly Rnor2.1, that are apparently functional, full-length genes in Rnor3.1. Consequently, some of the current 14 candidates, in particular those that are involved in fundamental processes of eukaryotic biology, may yet be 'repaired' by sequence changes in future assemblies, and thus be recognized as genic. However, genes associated with innate immunity (which is particularly susceptible to change via adaptive evolution), such as Forssman glycolipid synthetase and complement factor I, may yet be found to survive as true pseudogenes in the rat.

Non-coding RNA genes

We investigated the abundance and distribution of non-coding (nc)RNAs in rat. Cytoplasmic transfer (t)RNA gene identification in rodents is complicated by tRNA-derived identifier (ID) short interspersed nucleotide (SINEs) (B2 and ID). tRNAscan-SE predicted 175,943 tRNAs (genes and pseudogenes); however, the majority (175,285) were SINEs identified by RepeatMasker. This is far greater than the number found in mouse (24,402/25,078) or human (25/636). Of the remaining 666 predictions, 163 were annotated as tRNA pseudogenes and four were annotated as undetermined by tRNAscan-SE. An additional 68 predictions were removed because their best database match in either human, mouse or rat tRNA databases matched tRNAs with either a different amino acid or anticodon (violating the wobble rules that specify the distinct anticodons expected). The total of 431 tRNAs (including a single selenocysteine tRNA) identified in the rat genome is comparable to that for mouse—435 tRNAs (version mm2 from the UCSC genome browser)—and human—492 tRNAs (from the genomic tRNA database, http://rna.wustl.edu/GtRDB/Hs/). These three species share a core set of approximately 300 tRNAs, using a cutoff of ≥95% sequence identity and ≥95% sequence length.

A total of 454 ncRNAs (other than tRNAs) were identified by sequence comparison to known ncRNAs (Supplementary Information). These include 113 micro- (mi)RNAs, five ribosomal RNAs, 287 small nucleolar (sno)RNAs and small nuclear (sn)RNAs, 49 various other ncRNAs such as signal recognition particle (SRP) RNA, 7SK RNA, telomerase RNA, RNase P RNA, brain-specific repetitive (bsr)RNA, non-coding transcript abundantly expressed in brain (ntab)RNA, small cytoplasmic (sc)RNA and 626 pseudogenes. Complete 18S and 28S rRNA genes and more rRNAs were not identified, presumably owing to assembly issues.

Table 6 **Candidate rat pseudogenes, orthologous to mouse and human functional genes**

Mouse gene	Human gene	Strand	Rat genome coordinates*	Frameshifts/stops†	Annotation
ENSMUSG00000013611	ENSG00000174226	+	7:92752590–92807556	1/0	Sorting nexin
ENSMUSG00000024364	ENSG00000158402	+	18:62742414–62770427	2/0	Dual-specificity phosphatase CDC25c
ENSMUSG00000026293	ENSG00000077044	+	9:95634847–95692601	1/0	Diacylglycerol kinase δ
ENSMUSG00000026785	ENSG00000160447	+	3:9210762–9229984	5/0	Protein kinase PKNβ
ENSMUSG00000026829	ENSG00000148288	+	3:7662414–7664521	2/2	Forssman glycolipid synthetase
ENSMUSG00000027426	ENSG00000125846	+	3:125918806–125924149	1/1	Zinc finger protein 133
ENSMUSG00000028000	ENSG00000138799	−	2:221272797–221304350	1/0	Complement factor I
ENSMUSG00000029203	ENSG00000078140	−	14:44385206–44441888	1/0	Ubiquitin-protein ligase E2 (HIP2)
ENSMUSG00000030270	ENSG00000144550	−	20:8332585–8362331	3/0	Copine (membrane trafficking)
ENSMUSG00000035449	ENSG00000167646	+	1:67374986–67381472	1/0	Cardiac troponin I
ENSMUSG00000037029	ENSG00000105261	−	1:82728049–82730272	1/0	Zinc finger protein 146
ENSMUSG00000037432	ENSG00000158142	+	9:42465695–42498651	1/1	Dysferlin-like protein
ENSMUSG00000039660	ENSG00000167137	−	3:9320401–9326997	4/0	Similar to yeast YMR310c RNA-binding protein
ENSMUSG00000042653	ENSG00000137634	+	8:49938446–49939091	1/0	Brush border 61.9 kDa-like protein

*Coordinates from rat v2.0.
† Mouse genes were used as templates for predicting rat pseudogenes.

Evolution of transposable elements

Most interspersed repeats are immobilized copies of transposable elements that have accrued substitutions in proportion to their time spent fixed in the genome (for introduction[2,3,164–167]). About 40% of the rat genome draft is identified as interspersed repetitive DNA derived from transposable elements, similar to that for the mouse[3] (Table 7) and lower than for the human (almost 50%[2]). The latter difference is mainly due to the lower substitution rate in the human lineage, which allows us to recognize much older (Mesozoic) sequences as interspersed repeats. Almost all repeats are derived from retroposons, elements that procreate via reverse transcription of their transcripts. As in mouse, there is no evidence for activity of DNA transposons since the rat–mouse split. Many aspects of the rat and the mouse genomes' repeat structure are shared; here we focus on the differences.

LINE-1 activity in the rat lineage

The long interspersed nucleotide element (LINE)-1 (L1) is an autonomous retroelement, containing an internal RNA polymerase II promoter and two open reading frames (ORFs). The ORF1 product is an RNA binding protein with chaperone-like activity, suggesting a role in mediating nucleic acid strand transfer steps during L1 reverse transcription[168], whereas ORF2 encodes a protein with both reverse transcriptase and DNA endonuclease activity. LINEs are characteristically 5′ truncated so that only a small subset extends to include the promoter region and can function as a source for more copies.

Many classes of LINE-like elements exist, but only L1 has been active in rodents. Over half a million copies, in variable stages of decay, comprise 22% of the rat genome. Although 10% of the human genome is comprised of L1 copies introduced before the rodent–primate split, owing to the fast substitution rate in the rodent lineage only 2% of the rat genome could be recognized as such. Thus, probably well over one-quarter of all rat DNA is derived directly from the L1 gene.

Following the mouse–rat split, L1 activity appears to have increased in rat. The 3′ UTR sequences defined six rat-specific L1 subfamilies, represented by 150,000 copies that cover 12% of the rat genome. L1 copies accumulated over the same period in mouse cover only 10% of the genome (Table 7). This higher accumulation of L1 copies could explain some of the size difference of the rat and mouse genome.

In addition to the traditional L1 elements, there are 7,500 copies

Table 7 **Composition of interspersed repeats in the rat genome**

	Rat				Mouse	
	Copies ($\times 10^3$)	Total length (Mb)	Fraction of genome (%)	Lineage-specific (%)	Fraction of genome (%)	Lineage-specific (%)
LINEs	657	594.0	23.11	11.70	20.10	9.74
LINE-1	597	584.2	22.73	11.70	19.65	9.74
LINE-2	48	8.4	0.33	–	0.38	–
L3/CR1	11	1.4	0.06	–	0.06	–
SINEs	1,360	181.3	7.05	1.52	7.78	1.80
B1(Alu)	384	42.3	1.65	0.16	2.53	0.92
B4(ID_B1)	359	55.4	2.15	0.00	2.25	0.00
ID	225	19.6	0.76	0.54	0.20	0.00
B2	328	55.2	2.15	0.68	2.29	0.74
MIR	109	13.0	0.51	–	0.56	–
LTR elements	556	232.4	9.04	1.84	10.28	2.85
ERV_class I	40	24.9	0.97	0.56	0.79	0.36
ERV_class II	141	83.4	3.24	1.02	4.13	1.73
ERVL (III)	74	21.6	0.84	0.04	1.08	0.23
MaLRs	302	102.5	3.99	0.22	4.27	0.53
DNA elements	108	20.9	0.81	–	0.86	–
Charlie(hAT)	80	14.8	0.58	–	0.60	–
Tigger(Tc1)	18	4.0	0.16	–	0.17	–
Unclassified	14	7.3	0.28	–	0.37	–
Total	2,690	1,036	40.31	14.90	39.45	14.26
Small RNAs	8	0.6	0.03	0.01	0.03	0.01
Satellites	14	6.4	0.25	?	0.31	?
Simple repeats	897	61.1	2.38	?	2.41	?

Data for Rnor3.1 and October 2003 mouse (MM4), excluding Y chromosome, using the 17 December 2003 version of RepeatMasker. To highlight the differences between rat and mouse repeat content, columns 5 and 7 show the fractions of the genomes comprising lineage-specific repeats. The LINE-1 numbers include all HAL1 copies, whereas all BC1 scRNA and >10% diverged tRNA-Ala matches, far more common than other small RNA pseudogenes and closely related to ID, have been counted as ID matches.

(10 Mb) of a non-autonomous element that is derived from L1 by deletion of most of its ORF2. A similar element, active in Mesozoic times, has been called HAL1 (for Half-a-LINE)[164]. Given their low divergence, we conclude that the currently identified HAL1-like elements operated only a few million years ago in the mouse lineage (MusHAL1) and still propagate in the rat genome (RNHAL1). RNHAL1 contains only an ORF1, whereas MusHAL1 encoded an endonuclease as well, although no reverse transcriptase. The 5′ 2,600 bases of RNHAL1 are 98% identical to the currently active L1 in rat (L1_Rn or L1mlvi2[169]). Unlike ancient HAL1 elements, which shared the 3′ UTR with a contemporary L1, the 3′ end of RNHAL1 is unrelated to other repeats. The repeated origin and high copy number of HAL1s suggest that the ORF1 product, which binds strongly to its messenger RNA[168], may render this transcript a superior target for L1-mediated reverse transcription. In this way HAL1 resembles the non-autonomous, endogenous retrovirus-derived MaLR elements (below), which, for over 100 million years, retained only the retroviral gag ORF that encodes an RNA binding protein. A potential advantage of HAL1 over L1 is its shorter length, which, considering the usual 5′ truncation of copies, increases the chance that a copy may include the internal promoter elements and become a source gene.

Different activity of SINEs in the rat and mouse lineage

The most successful usurpers of the L1 retrotransposition machinery, however, are SINEs. These are small RNA-derived sequences with an internal RNA polymerase III promoter. Recently, the human Alu SINE has been experimentally proven to be transposed by L1[170]. Most SINEs share the 3′ end with their associated LINE elements, like the Mesozoic mammalian LINE-2 (L2) and MIR pair, increasing the efficiency with which a LINE reverse transcriptase recognizes the 3′ end of a dependent SINE. However, L1 does not show sequence specificity and rodent and primate SINE sequences are unrelated to L1. Although any transcript can be retroposed, as can be seen from the numerous processed pseudogenes in mammalian genomes, L1-dependent SINEs probably have features that make them especially efficient targets of the L1 reverse transcriptase.

Although before the radiation of most mammalian orders L1 was at least as active as L2, the L2-dependent MIR was the only known (and very abundant) SINE of that time. All of the currently active SINEs in different mammalian orders appear to have arisen after the demise of L2 (and consequently MIR), as though an opportunity (or necessity) arose for the creation and expansion of other SINEs.

Four different SINEs are distinguished in rat and mouse. The B1 element seems to share its origin from a 7SL RNA gene with the primate Alu[171]. This probably happened just before the rodent–primate split and after the speciation from most other eutherians, where Alu/B1 elements are not known. The other SINEs are rodent-specific and have tRNA-like internal promoter regions. ID elements consist only of this tRNA-like region, which in older ID copies closely match an Ala-tRNA from which it may have been derived. B4 resembles a fusion of an ID and B1 SINE. Finally, B2 has a tRNA-like region of unknown affiliation followed by a unique 120 bp region.

The fortunes of these SINEs during mouse and rat evolution have been different (Fig. 12). B4 probably became extinct before the mouse–rat speciation, while B2 has remained productive in both lineages, scattering >100,000 copies in each genome after this time. Interestingly, the fate of the B1 and ID SINEs has been opposite in rat and mouse. While B1 is still active in mouse, having left over 200,000 mouse-specific copies in its trail, the youngest of the 40,000 rat-specific B1 copies are 6–7% diverged from their source, indicating a relatively early extinction in the rat lineage. On the other hand, after the mouse–rat split only a few hundred ID copies may have inserted in mouse, whereas this previously minor SINE (~60,000 copies predate the speciation) increased its activity in rat to produce 160,000 ID copies.

Co-localization of SINEs in rat and mouse

Despite the different fates of SINE families, the number of SINEs inserted after speciation in each lineage is remarkably similar: ~300,000 copies. Reminiscent of the replacement of MIR by L1 driven SINEs, it seems that the demise of B1 in rat allowed the expansion of IDs. Moreover, these independently inserted and unrelated SINEs (ID and B1 share only a mechanism of retroposition) accumulated at orthologous sites: the density of rat-specific SINEs in 14,243 ~100 kb windows in the rat genome is highly correlated ($R^2 = 0.83$) with the density of mouse-specific SINEs in orthologous regions in mouse. To avoid including elements fixed before the speciation, only SINEs labelled lineage-specific on the basis of subfamily assignment (Methods[89]) were tallied with a divergence from the consensus that was well below the 9% average for neutral sites (Fig. 5). These data corroborate and refine the observation of a strong correlation between the location of primate- and rodent-specific SINEs in 1 Mb windows[3]. At 100 kb, no correlation is seen for interspersed repeats other than SINEs.

Insertions of SINEs at the same location in different species have been reported[172–174], and the correlation could reflect the existence of conserved hotspots for SINE insertions. However, only five of ~800 human specific Alu elements have an Alu inserted within 100–200 bp in any of six other primate lineages[174–176]. Likewise, gene conversions of shared Alus into lineage-specific copies were observed five times in the same set, too low a level to contribute significantly to the observed correlation[174–176].

Figure 9c displays the lineage-specific SINE densities on rat chromosome 10 and in the mouse orthologous blocks, showing a stronger correlation than any other feature. The cause of the unusual distribution patterns of SINEs, accumulating in gene-rich regions where other interspersed repeats are scarce, is apparently a conserved feature, independent of the primary sequence of the SINE and effective over regions smaller than isochores.

In the human genome, the most recent (unfixed) Alus are distributed similarly to L1, whereas older copies gradually take on the opposite distribution of SINEs[2,164]. This suggested that SINEs insert in the same places as LINEs, and that the typical SINE pattern is due to selection (or deletion bias) rather than a mechanistic insertion bias shared by all (unrelated) SINEs, but not by LINEs that use the same insertion process. This led to a proposal that SINEs are preferentially maintained in regions where they can easily be expressed[2,164]: if so, this could be the local feature conserved between mammalian genomes that leads to the strong correlation of local SINE densities in different mammals. However, we did not observe this temporal shift in SINE distribution pattern in mouse, nor currently in the rat genome, despite a considerable effort to define the potentially unfixed SINEs in both species (see ref. 89 for details). The observations in human could reflect a recent change in Alu behaviour, which would necessitate another explanation for the contrary insertion-preference of older Alus and all other SINEs.

Some regions of high LINE content coincide with regions that exhibit both higher AT content and an increased rate of point substitution (Fig. 9, pink rectangles). In a genome-wide analysis, LINE content correlates strongly with substitution rates, and about 80% of this correlation is explained by higher rates in AT-rich regions[89]. SINE density shows the opposite correlation both on chromosome 10 (Fig. 9) and genome-wide[89].

These phenomena, in conjunction with an overall trend in substitution rates towards AT-richness, suggest a model in which quickly evolving regions accumulate a higher-than-average AT content, which attracts LINE elements. Although distinct cause–effect relationships such as this remain largely speculative, these results reinforce the idea that local genomic context strongly shapes local genomic features and rates of evolution.

Endogenous retroviruses and derivatives

The other major contributors to interspersed repeats in the rodent

genome are retrovirus-like elements. These have several 100 bp long terminal repeats (LTRs) with transcriptional regulatory sequences that flank an internal sequence that, in autonomous elements, encodes all proteins necessary for retrotransposition. All mammalian LTR elements are endogenous retroviruses (ERVs) or their non-autonomous derivatives. They fall into three groups, of which representatives in mouse are: murine leukaemia virus (MuLV) (class I), intracisternal A-particle (IAP) and MMTV (class II), and MERVL (class III).

The most productive retrovirus in mammals has been the class III element ERV-L, primarily through its ancient non-autonomous derivatives, called MaLRs, with 350,000 copies occupying ~5% of the rat genome (Table 7). Human ERV-L and MaLR copies are >6% diverged from their reconstructed source genes and must have died out around the time of human speciation from New World monkeys. In mouse, several thousand almost identical MaLR and ERV-L copies suggest sustained activity[177–179]. In contrast, rat ERV-L activity must have been silenced a few million years ago, given that the least diverged MaLR and ERV-L (MTB_Rn and MT2_Rat1) copies differ by >4% from each other. Other class III ERVs were active earlier in rodent evolution, before the mouse–rat speciation.

In contrast to class III ERVs, class I and class II elements still thrive in rat. We reconstructed four rat-specific autonomous class I ERVs, of which two appear still active, and nine class II ERVs, of which four may still be active. The non-autonomous NICER and RAL elements represent over 60% of all rat-specific class I elements. The autonomous drivers of this group, RNNICER2 and 3, with several intact copies, are closely related to the mouse-specific MuLV. Among the potentially active autonomous class II ERVs are MYSERV_Rn, related to the Mys element in *Peromyscus*, and several IAP elements, one with a full-length envelope gene. The most prolific, still-active class II ERV, RNERVK3, is distantly related to the simian retroviruses and, like ERV-L and NICER, has spawned abundant non-autonomous elements characterized by closely related LTRs.

Simple repeats

Whereas the above interspersed repeats derive from transposed sequences, mammalian genomes also contain interspersed simple sequence repeats (SSRs), regions of tandemly repeated short (1–6 bp) units that probably arise from slippage during DNA replication and can expand and compress by unequal crossing over. Remarkable differences were noted between the SSR contents of the human and mouse genomes[3]. Three times as many base pairs are contained in near (>90%) perfect SSRs in mouse than in human, and a 4–5-fold excess was revealed when excluding SSRs contained in or seeded by interspersed repeats (primarily SSRs derived from the poly A or simple repeat tails of SINEs and LINEs). SSRs are both more frequent and on average longer in mouse. Polypurine (or polypyrimidine) repeats are especially (tenfold) over-represented in the mouse genome. As discussed above, this contrasts sharply with the greater frequency of triplet repeats coding for amino acids in human than in the rodents.

Rat and mouse SSR contents show, perhaps not surprisingly, much smaller differences. They represent almost the same amount of the rat and mouse genomes (for >90% perfect elements, ~1.4% compared with 0.45% in human) and are of similar average length; for example, the average >90% perfect $(CA)_n$ repeat, the most common SSR in mammals, is 42 bp long in mouse and 44 bp in rat. Some potentially significant differences are that polypurine SSRs are of similar average length but are 1.2-fold more common in mouse, whereas the rare SSRs containing CG dimers are 1.5-fold more frequently observed in rat.

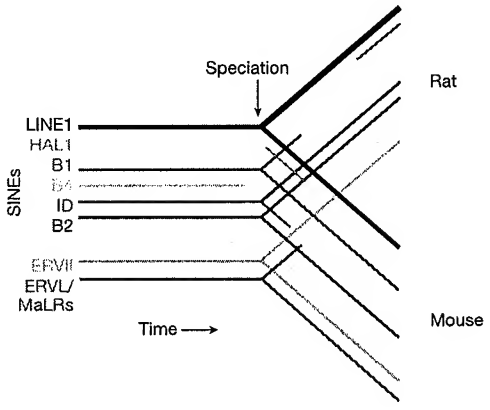


Figure 12 Historical view of rodent repeated sequences. Relationships of the major families of interspersed repeats (Table 7) are shown for the rat and mouse genomes, indicating losses and gains of repeat families after speciation. The lines indicate activity as a function of time. Note that HAL1-like elements appear to have arisen in both the mouse and rat lineages.

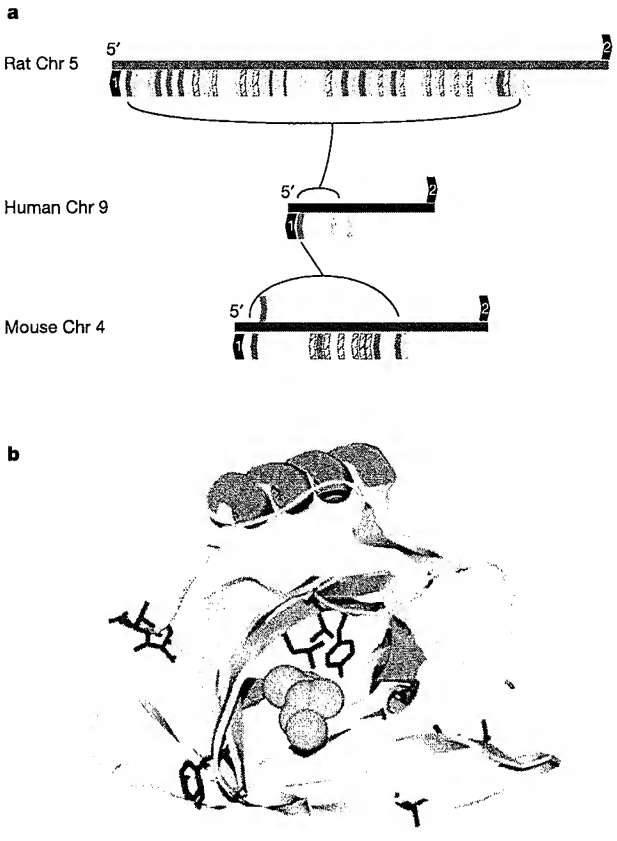


Figure 13 Adaptive remodelling of genomes and genes. **a**, Orthologous regions of rat, human and mouse genomes encoding pheromone-carrier proteins of the lipocalin family (α_{2u}-globulins in rat and major urinary proteins in mouse) shown in brown. Zfp37-like zinc finger genes are shown in blue. Filled arrows represent likely genes, whereas striped arrows represent likely pseudogenes. Gene expansions are bracketed. Arrowhead orientation represents transcriptional direction. Flanking genes 1 and 2 are *TSCOT* and *CTR1*, respectively. **b**, Site-specific K_A/K_S analysis of rat α_{2u}-globulins. Shown in red are side-chains from codons subject to positive selection. These have been mapped to a ribbon representation of the crystal structure of rat α_{2u}-globulin chain A.

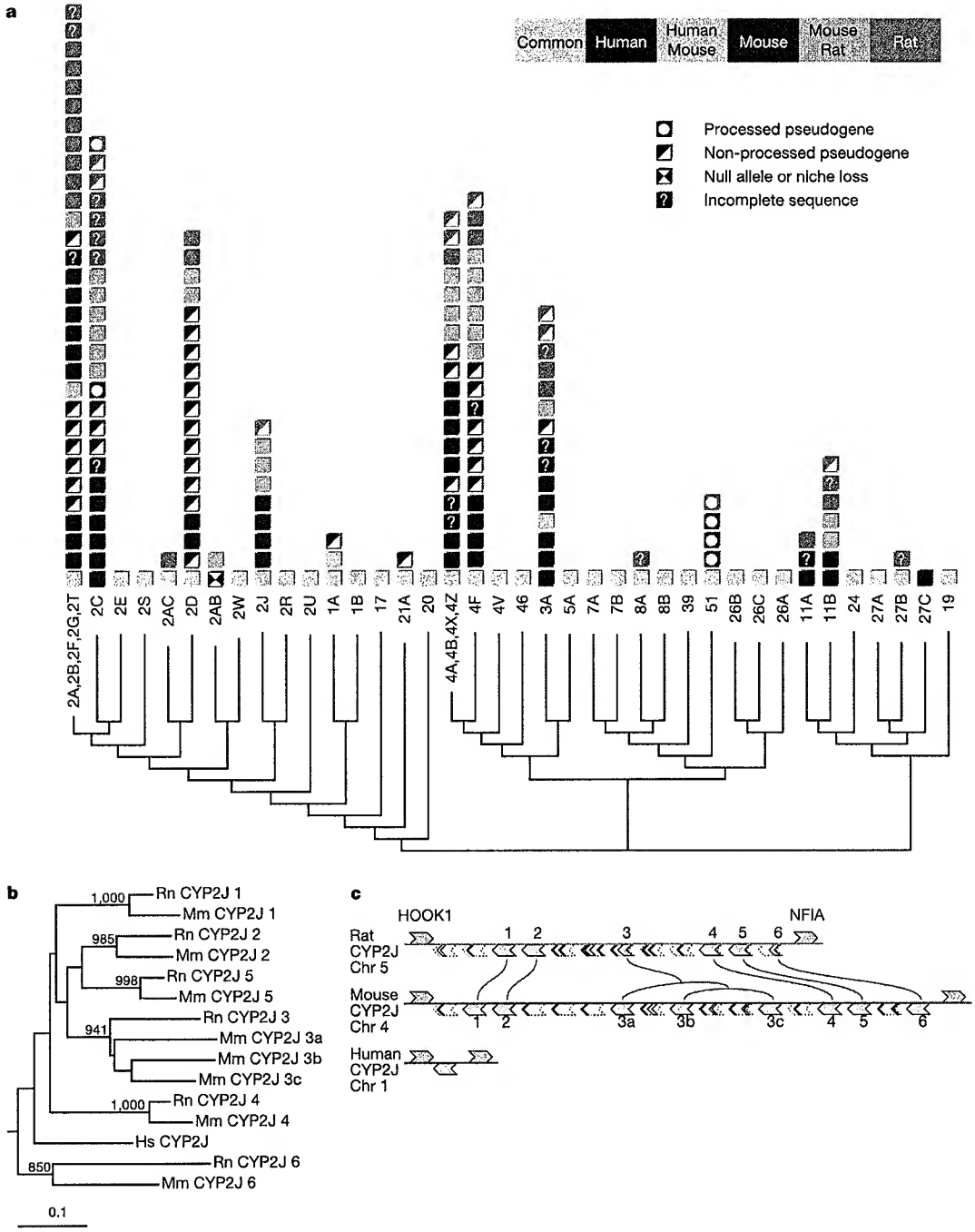


Figure 14 Evolution of cytochrome P450 (CYP) protein families in rat, mouse and human. **a**, Dendrogram topology from 234 full-length sequences. 279 sequences of ≥300 amino acids; subfamily names and chromosome numbers are shown. Black branches have >70% bootstrap support. Incomplete sequences (they contain Ns) are included in counts of functional genes (84 rat, 87 mouse and 57 human) and pseudogenes (including fragments not shown; 77 rat, 121 mouse and 52 human). 64 rat genes and 12 pseudogenes were in predicted gene sets. Human CYP4F is a null allele owing to an in-frame STOP codon in the genome, although a full-length translation exists (SwissProt P98187). Rat CYP27B, missing in the genome, is 'incomplete' because there is a RefSeq entry (NP_446215). Grouped subfamilies CYP2A, 2B, 2F, 2G, 2T and CYP4A, 4B, 4X, 4Z, occur in gene clusters; thus nine loci contain multiple functional genes in a species. One (CYP1A) has fewer rat genes than human, seven have more rodent than human, and all nine differ in rodent copy numbers. CYP2AC is a rat-specific subfamily (orthologues are pseudogenes). CYP27C has no rodent counterpart. Rodent-specific expansion, rat CYP2J, is illustrated below. **b**, The neighbour-joining tree[224], with the single human gene, contains clear mouse (Mm) and rat (Rn) orthologous pairs (bootstrap values >700/1,000 trials shown). Bar indicates 0.1 substitutions per site. **c**, All rat genes have a single mouse counterpart except for CYP2J 3, which has further expanded in mouse (mouse CYP2J 3a, 3b and 3c) by two consecutive single duplications. The genes flanking the CYP2J orthologous regions (rat chromosome 5, 126.9–127.3 Mb; mouse chromosome 4, 94.0–94.6 Mb; human chromosome 1, 54.7–54.8 Mb) are hook1 (HOOK1; pink) and nuclear factor I/A (NFIA; cyan). Genes (solid) and gene fragments (dashed boxes) are shown above (forward strand) and below (reverse strand) the horizontal line. No orthology relation could be concluded for most of these cases.

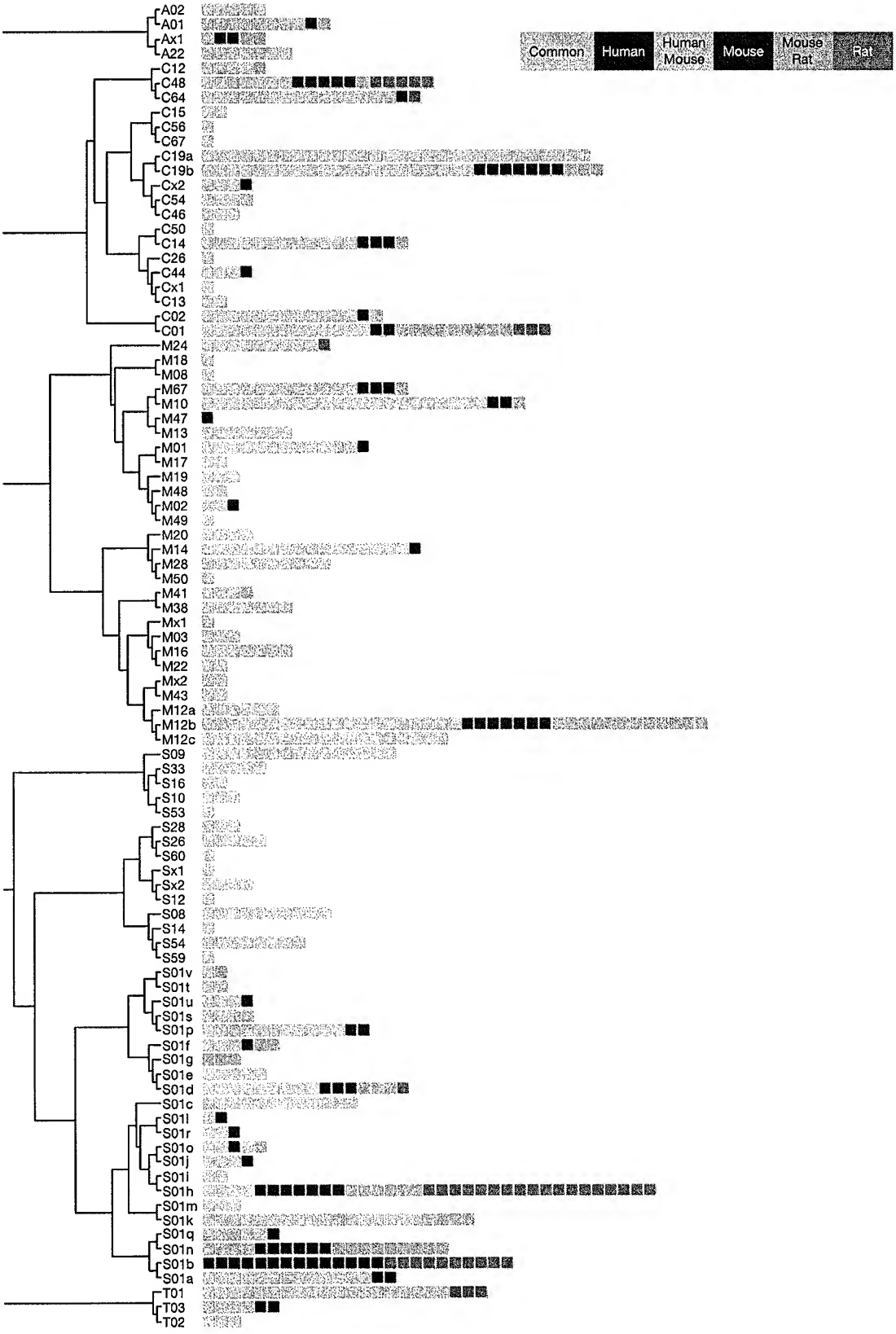


Figure 15 Comparative analysis of rat, mouse and human proteases. The complete non-redundant set of proteases and protease homologues from each species is distributed in five catalytic classes and 67 families. Each square represents a single protease, and is coloured according to its presence or absence in rat, mouse and human as indicated in the inset.

Prevalent, medium-length duplications in rodents

In addition to the transpositionally derived interspersed repeats and simple repeats detected by RepeatMasker and Tandem Repeat Finder, the rat and mouse genomes contain a substantial amount of medium-length unclassified duplications (typically 100–5,000 bp). These are readily seen in self-comparisons and in intra-rodent comparisons after masking the known repeats, but they are substantially less prevalent in comparisons with the human genome (Supplementary Information). Clearly, a substantial fraction of the rodent genomes consists of currently unexplained repeats and a full characterization awaits further studies. The unclassified duplications may include: (1) novel families of low-copy rodent interspersed repeats; (2) extensions of known but not fully characterized rodent repeats; and (3) duplications generated by a mechanism different from transposition.

Rat-specific biology

A principal ambition of the RGSP was to reveal genetic differences between rats and mice that might specify their differences in physiology and behaviour. This view was well supported by the current draft sequence and predicted gene set. In particular, recently duplicated genes are enriched in elements involved in chemosensation and functional aspects of reproduction (Table 5). Here we illustrate the differences in the gene complements of rat and mouse by in-depth analyses of olfactory receptors (ORs), pheromones, cytochromes P450, proteases and protease inhibitors.

Chemosensation

The ability to emit and sense specific smells is a key feature of survival for most animals in the wild. Another paper[180] describes the evolution of rat and mouse pheromones, vomeronasal receptors, and ORs whose genes were duplicated frequently during the time since the common ancestor of rats and mice (Table 5). Their study yielded over 200 aligned codons predicted to have been subject to adaptive evolution. They attribute the rapid evolution of these genes to conspecific competition—in particular, sexual selection.

Using a homology-based identification procedure with manual curation[181], we found 1,866 ORs in 113 locations in the rat genome: 69 multi-gene clusters and 44 single genes. After adjusting for missing sequences (the assembly covers 90.2% of the genome), we extrapolate that there are ~2,070 OR genes and pseudogenes. The rat therefore has ~37% more OR genes and pseudogenes than the ~1,510 ORs of the mouse[181,182], assuming similar representation of recently duplicated sequences in the two genome assemblies used. Of the 1,774 OR sequences that are not interrupted by assembly gaps, 1,227 (69%) encode intact proteins, while the remaining 547 (31%) sequences are probably pseudogenes with in-frame stop codons, frameshifts, and/or interspersed repeat elements. Fewer mouse OR homologues are pseudogenes (~20%)[181,182], but the larger family size in rat still leaves it with substantially more intact ORs than the mouse (~1,430 versus ~1,210). Striking rat-specific expansions of two ancestral clusters account for much of the difference in OR family size and pseudogene content between rat and mouse, although many other clusters exhibit more subtle changes (not shown). Significant differences between human and mouse OR families have also been reported[181–183], but the functional implications of OR repertoire size on the ability of different species to detect and discriminate odorants are not yet known.

α_{2u}-globulin pheromones

The α_{2u}-globulin genes are odorant-binding proteins that also contribute to essential survival functions in animals. α_{2u}-globulin homologues are likely to be highly heterogeneous among murid species. Several homologues (major urinary proteins) sequenced from the BALB/c mouse are distinct from their C57BL/6J mouse counterparts, and these also appear to be arranged differently along its genome[184]. Moreover, two full-length genes from other mouse strains[185] differ from their C57BL/6J orthologues—either lacking two of the bases or retaining 20 of the bases that render the C57BL/6J sequences likely to be pseudogenes (not shown).

The evolution of α_{2u}-globulin genes on rat chromosome 5 has clearly driven a significant 'remodelling' of this genomic region (Fig. 13a). The orthologous human genomic region contains a single homologue, suggesting that the common ancestor of rodents and human possessed one gene. The genome of C57BL/6J mice contains four homologous genes, and seven pseudogenes, whereas the rat genome contains ten α_{2u}-globulin genes and 12 pseudogenes in a single region (Fig. 13a).

Phylogenetic trees constructed using amino acid, and non-coding DNA, sequences show that, surprisingly, the rat α_{2u}-globulin gene clusters appear to have arisen recently via a rapid burst of gene duplication since the rat–mouse split (Table 5; data not shown). This is consistent with the Rfp37-like zinc-finger-like pseudogene having uniquely 'hitchhiked' for virtually all of the rat-specific α_{2u}-globulin gene duplications (Fig. 13a). The sequences of these genes are also evolving rapidly, with median K_A/K_S values of 0.77 and 1.06 for rat and mouse genes, respectively. Amino acid sites that appear to have been subject to adaptive evolution are situated both within the ligand-binding cavity, and on the solvent-exposed periphery of the α_{2u}-globulin structure[139] (Fig. 13b). This demonstrates how genome analysis can reveal the imprint of adaptive evolution from megabase to single-base levels.

The rapid evolution of these genes, and the remodelling of their genomic regions, can be attributed to the known roles of rat α_{2u}-globulins and mouse major urinary proteins in conspecific competition and sexual selection. These proteins are pheromones and pheromone carriers that are present in large quantities in rodent urine, and act as scent markers indicating dominance and subspecies identity[186,187].

Detoxification

Cytochrome P450 is a well-recognized participant in metabolic detoxification, and we also observe rapid evolution within this family. These enzymes metabolize a large number of toxic and endogenous compounds[188] and thus are particularly relevant to clinical and pharmacological studies in humans. As rodents are important model organisms for understanding human drug metabolism, it is important to identify 1:1 orthologues and species-specific expansions and losses[189]. Compared with human genes, there are clear expansions of several rodent P450 subfamilies, but there are also significant differences between rat and mouse subfamilies (Fig. 14a). The fastest-evolving subfamily seems to be CYP2J, containing a single gene in human, but at least four in rat and eight in mouse (Fig. 14b, c). CYP2J enzymes catalyse the NADPH-dependent oxidation of arachidonic acid to various eicosanoids, which in turn possess numerous biological activities including modulation of ion transport, control of bronchial and vascular smooth muscle tone, and stimulation of peptide hormone secretion[190]. The genomic ordering of genes and their phylogenetic tree indicate an ongoing expansion in the rodents (Fig. 14b, c). This suggests that adaptive evolution has been involved in diversifying their functions. Moreover, detailed study of the nuclear receptors, a highly conserved family of transcription factors, revealed that PXR and CAR, two nuclear receptors regulating CYP genes involved with detoxification[191], have the two highest nucleotide substitution rates in their ligand binding domains, whereas SF-1, the nuclear receptor regulating CYP19 (ref. 192), which has not undergone expansion, is more conserved, like other nuclear receptors[193].

Proteolysis

Protease and protease inhibitor genes also represent an example of rapid evolution in the rat genome. Proteases are a structurally and functionally heterogeneous group of enzymes involved in multiple biological and pathological processes[194]. The rat contains 626

protease genes, ~1.7% of the rat gene count[124], more than human (561) but similar to mouse (641)[125]. Of the rat protease genes, 102 are absent from human, and 42 are absent from mouse (Fig. 15). Several rat gene families have expanded, including placental cathepsins, testases, kallikreins and haematopoietic serine proteases; others appear to have formed pseudogenes in humans (Table 8). These protease families are mainly involved in reproductive or immunological functions, and have evolved independently in the rat and mouse lineages.

The rat protease inhibitor complement contains 183 members, similar to mouse (199) but larger than human (156). As with the protease genes, the rapid evolution in protease inhibitors derives from differential expansions of specific families such as serpins and cystatins. The concomitant expansions in rat and mouse proteases and their inhibitors appear to reflect homeostasis of protein turnover.

These gene family expansions dramatically illustrate how large-scale genomic changes have accompanied species-specific innovation. Positive selection of duplicated genes has afforded the rat an enhanced repertoire of precisely those genes that allow reproductive success despite severe competition from both within its own, and with other, species. This serves as a general illustration of the importance of chemosensation, detoxification and proteolysis in innovation and adaptation.

Human disease gene orthologues in the rat genome

A further strong motivation for sequencing the rat genome was to enhance its utility in biomedical research. Although the rat is already recognized as the premier model for studying the physiological aspects of many human diseases, it has not had as prominent a role in the study of simple genetic disease traits. As more than 1,000 human mendelian disorders now have associated loci and alleles, there is now a tremendous opportunity to link the new knowledge of the rat genome with data from the human disease examples. The precise identification of the rat orthologues of human genes that are mutated in disease creates further opportunities to discover and develop rat models.

Predicted rat genes were compared with 1,112 well-characterized human disease genes[195] that were verified and classified on the basis of pathophysiology (H.H., E.E.W., H.W., K.G.W., H.X., L.G., P.D.S., D.N.C., D.S., M.M.A., C.P.P. and K.F., unpublished work). As predicted by Ensembl, 844 (76%) have 1:1 orthologues in the rat. These predictions are likely to be of high quality because 97.4% of the 11,422 rat:human 1:1 orthologues predicted by Ensembl were found in orthologous genomic regions.

We asked if these 'disease orthologue' pairs were distinguishable from other rat–human orthologues. Ensembl automatically predicts that 11,522 human genes have rat 1:1 orthologues (corresponding to 46% of all Ensembl predicted human genes). By contrast, a much higher proportion (76%) of human disease genes have Ensembl-predicted rat 1:1 orthologues. Careful analysis of the remaining 268 human genes that were not predicted by Ensembl to show 1:1 orthology indicated that only six of the human disease genes lack likely rat orthologues among genome, cDNA, EST and protein sequences[196]. Thus, it appears that, in general, genes involved in human disease are unlikely to have diverged, or to have become duplicated, deleted or lost as pseudogenes, between rat and human (conservation of orthologues discussed above).

We next compared K_S, K_A and the K_A/K_S ratio values of 'disease orthologues' with those of all remaining orthologue pairs. Only the K_S distributions differed significantly[196], suggesting that coding regions of human disease genes and their rat counterparts have mutated more rapidly than the non-disease genes. This might result from factors influencing the specific loci, or the disease genes may characteristically reside in genomic regions that exhibit higher mutation rates.

The disease gene set was next grouped into 16 disease-system categories and analysed using a non-parametric test for K_A/K_S (human/rat)[196] (Fig. 16). Only five disease systems exhibited significant K_A/K_S differences with respect to the remaining samples ($P < 0.05$). Neurological and malformation-syndrome disease categories manifested the lowest median K_A/K_S ratios that are consistent with purifying selection acting on these gene sets. With a comparison of the mean to the mean and standard deviation of the null hypothesis, [(Mean–Mean0)/Std0] of −4.63 ($P < 0.0001$), the neurological disease gene set revealed the most evidence for purifying selection of the disease gene categories examined. In contrast, the pulmonary, haematological and immune categories manifested the highest median K_A/K_S ratios, and the genes of the immune system disease category, with a value for (Mean–Mean0)/Std0 of 4.98 ($P < 0.0001$), show the highest K_A/K_S ratios. These results are consistent with a role for more positive selection, or reduced selective constraints, among these genes.

Where possible, we further considered conservation of these pathophysiology-based gene sets among orthologues of more diverse phyla, including mouse, fish, fly, nematode worm and

Table 8 **Protease-expanded gene families and pseudogenes in rat, mouse and human genomes**

Protease	Rat gene / locus	Human gene / locus	Mouse gene / locus	Function
Absent genes in assembly	13 from 626 (2.07%)	5 from 561 (0.89%)	5 from 641 (0.78%)	
Expanded families				
Placental cathepsins	10 genes / 17p14	Absent	8 genes / 13B3	Reproduction
Testins	3 genes / 17p14	Absent	3 genes / 13B3	Reproduction
Glandular kallikreins	10 genes / 1q21	Absent	15 genes / 7B2	Reproduction
Mast cell chymases/granzymes	28 genes / 15p13	4 genes / 14q11	17 genes / 14C1	Host defence
Human pseudogenes				
Chymosin	1 gene / 2q34	1 ps / 1p13	1 gene / 3F3	Digestion
Distal intestinal serine proteases	2 genes / 10q12	1 ps / 16p12	2 genes / 17A3	Digestion
Pancreatic elastase	1 gene / 7q35	1 ps / 12q13	1 gene / 15F3	Digestion
Fertilins and reproductive ADAMs	7 genes / various loci	6 ps / various loci	8 genes / various loci	Reproduction
Testases	4 genes / 16q12	3 ps / 8p22	9 genes / 8B1	Reproduction
Testis serine proteases	5 genes / various loci	5 ps / various loci	6 genes / various loci	Reproduction
Implantation serine proteases	2 genes / 10q12	1 ps / 16p13	2 genes / 17A3	Reproduction
Airway trypsin-like proteases	3 genes / 14p21	3 ps / 4q13	3 genes / 5E1	Host defence
Rat pseudogenes				
Calpain 13	1 ps / 6q12	1 gene / 2p23	1 gene / 17E2	Reproduction ?
Pyroglutamyl-peptidase II	1 ps / 1q22	1 gene / 15q26	1 gene / 7C	Metabolism
Gln-fructose-6-P transamidase 3	1 ps / Xq14	1 gene / Xq21	1 ps / XC3	Metabolism
Aminopeptidase MAMS/L-RAP	1 ps / 1q12	1 gene / 5q15	1 ps / 17A3	Host defence
Carboxypeptidase O	1 ps / 9q31	1 gene / 2q33	1 ps / 1C2	Unknown
Procollagen III *N*-endopeptidase	1 ps / 19q12	1 gene / 16q24	1 ps / 8E2	Metabolism ?
Kallikrein-2 and -3	2 ps / 1q21	2 genes / 19q13	1 ps / 7B2	Reproduction
Testis-specific protein 50	1 ps / 8q32	1 gene / 3p21	1 gene / 9F2	Reproduction

yeast orthologues. Overall, we obtained results consistent with those reported here for these rat:human 1:1 orthologous gene disease categories[196]. These results demonstrate that the individual genes that constitute various disease systems exhibit significantly different average evolutionary rates. The higher evolutionary rates noted for the immune system disease genes are consistent with a previous finding that lymphocyte-specific genes evolve relatively rapidly[197] and may indicate rapid diversification of the functions of the immune systems of rodents and humans. This is expected for genes involved in controlling species-restricted infectious agents if strong adaptive pressure acts during host–pathogen co-evolution. Thus, the results of studies of these rodent genes may be less directly relevant to our understanding of human immune system diseases than results obtained for other pathophysiology disease systems where conservation is greater and purifying selection is stronger.

We have also specifically examined a number of genes that harbour triplet nucleotide repeats, and are involved in human neurological disorders such as Huntington's disease, a condition known to be caused by CAG triplet repeat expansion producing abnormally long polyglutamine tracts in an otherwise normal protein[198]. Analysis of the rat–human orthologues of these disease genes indicated that repeat-expansion disease genes exhibit a repeat length that is substantially shorter in the rat than that found in the normal human gene (Fig. 17). In all cases, human disease genes localize below the line demarcating 1:1 length correlation, showing that rat orthologues uniformly bear shorter repeats. At present, there are no naturally occurring rat strains described that exhibit neurological disease associated with repeat-expansion mechanisms. The shorter repeat length of these orthologues in the rat would be consistent with either the lack of repeat-expansion mutational mechanisms in the rat or the failure of these orthologues to achieve a 'critical repeat length' susceptible to such mutational mechanisms. Other human genes, not at present known to be associated with disease, also contain glutamine repeats that are much shorter in the rat orthologues, and thus, could be investigated as potential disease candidates[196]. These triplet-repeat-bearing genes may be susceptible to mutations that arise through repeat-expansion mechanisms. In Fig. 17, it may also be observed that a relatively high proportion of repeats are significantly longer in the rat than in their corresponding human orthologue.

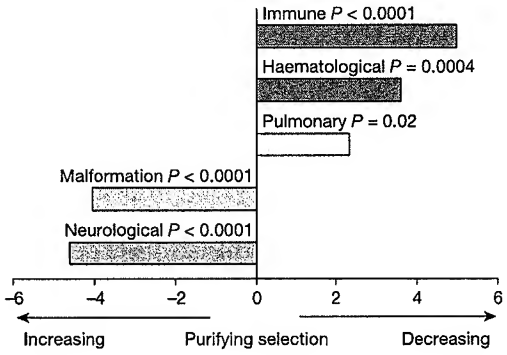


Figure 16 Selective constraints differ for human disease systems in the rat genome. Human disease system categories showing significant differences ($P < 0.05$) in a non-parametric test (Mann–Whitney–Wilcoxon) comparing K_A/K_S (human:rat) ratios. *P* values from two-level tests between genes from one disease system and the remaining genes. (Mean–Mean0)/Std0 values from multi-level tests from 16 categorized disease systems. Negative values (shown in yellow and orange) for neurological (−4.63) and malformation-syndrome (−4.04) categories were observed to be consistent with K_A/K_S ranges in which purifying selection predominates. Immune, haematological and pulmonary categories show positive values of 4.98, 3.59 and 2.34, respectively (for complete data set and details, see ref. 199).

In addition to enabling the direct comparison of rat–human disease orthologues, the rat genome sequence itself is an invaluable aid for the discovery of additional rat genes that can be studied as disease models. Two general modes can now be pursued. First, genes underlying disease phenotypes with simple inheritance that have been mapped to chromosomal regions can be more easily pursued in both species. Indeed, the rearrangements of conserved segments between the two species in this map were found to have significant value, because they tighten the boundaries of the mapped disease regions and thus reduce the number of genes that could potentially be associated with a given disease phenotype[113]. Second, the identification of multiple alleles contributing to quantitative and complex trait differences that are involved in disease processes can be pursued with more accuracy, both in the initial association phases, and in subsequent efforts to detect causative alleles.

Rat single nucleotide polymorphisms

The discovery and cataloguing of the natural DNA variation that persists between individual rat strains will allow further research using rat model systems. Although many rat microsatellites have been characterized and studied, single nucleotide polymorphisms (SNPs) are of more general interest because of their probable ubiquity, and the ease with which they can be assayed. SNP data have three broad applications: (1) the individual markers can be used in ongoing efforts to associate phenotypes that have complex underlying genetic components, with specific sites in the genome. (2) A panel of such markers can be used in conjunction with selective breeding and chromosome mechanics, to generate rat strains that are amenable to the kinds of manipulations that will hasten the discovery of important alleles. (3) A set of such markers can be used to detail the history of the different genomic events that have led to the structure of the genomes of contemporary rat strains. A detailed map of these events has a utility analogous to the current human haplotype (HapMap) mapping project[199] and will probably

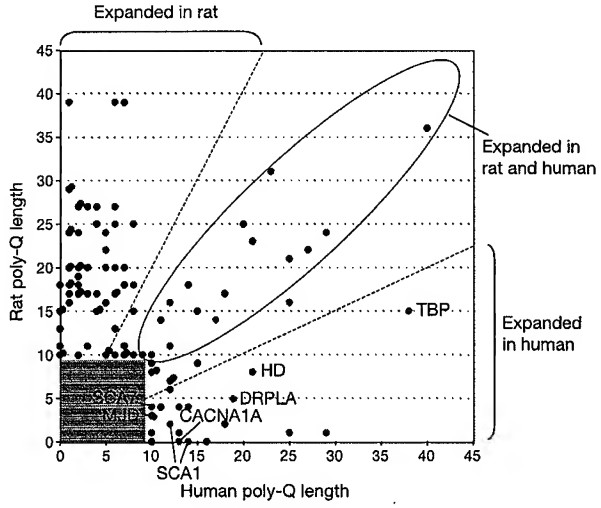


Figure 17 Polyglutamine repeat length comparison between human and rat. Points represent protein poly-Q length for rat and human. Red points correspond to repeats in genes associated with human disease: SCA1, spinocerebellar ataxia 1 protein, or ataxin1; SCA7, spinocerebellar ataxia 7 protein; MJD, Machado–Joseph disease protein; CACNA1A, spinocerebellar ataxia 6 protein, or calcium channel alpha 1A subunit isoform 1; DRPLA, dentatorubral pallidoluysian atrophy protein; HD, Huntington's disease protein, or huntingtin; TBP, TATA binding protein or spinocerebellar ataxia 17 protein. Repeat lengths over ten were examined; green shading delineates the range not included in our analysis. Also noted are a set that are expanded in rat and human (black circle) and a set where repeats are expanded in the rat.

aid disease gene identification, as recently suggested for the mouse[200].

The Rnor3.1 draft sequence was generated primarily from DNA of a single inbred rat line. This maximized the likelihood of deriving an accurate sequence assembly, but reduced any likely discovery of natural variation in this phase of the project. As a consequence there has been no large-scale public SNP discovery from rat genomic sequencing. A pilot project based on coding (c)SNP discovery has been initiated, however[201], as these cSNPs represent a particularly important subset of variants that may have direct functional significance[202]. These data have illustrated both immediate applications and the long-term potential for an effort aimed at comprehensive SNP discovery.

Conclusions

As the third mammalian genome to be sequenced, the rat genome has provided both predictable and surprising information about mammalian species. Although it was clear at the outset of this programme that ongoing rat research would benefit from the resource of a genome sequence, there was uncertainty about how many new insights would be found, especially considering the superficial similarities between the rat and the already sequenced mouse. Instead, the results of the sequencing and analysis have generated some deep insights into the evolutionary processes that have given rise to these different species. In addition, the project has been invaluable in further developing the methods for the generation and analysis of large genome sequence data sets.

The generation of the rat draft tested the new 'combined approach' for large genome sequencing. As the overall assembly is of high quality, there is no doubt that this overall strategy, and the supporting software we have developed, provides a suitable approach for this problem. Because we included a BAC 'skimming' component in the underlying data set, the assembly recovered a fraction of the genome that was expected, by analogy to the mouse project, to be difficult to assemble from pure WGS data. In addition, the BAC skimming component allowed progressive generation of high-quality local assemblies that were of use to the rat research community as the project developed. On the other hand, although the BAC component used here was far less expensive than the fully ordered and highly redundant set used in the hierarchical approach to sequencing the human genome, it nevertheless increased the overall cost of data production relative to a WGS approach.

The issue of efficacy of WGS versus other approaches to the sequencing of large genomes remains a matter of earnest scientific debate. In ongoing projects at different centres that participated in the RGSP consortium, different approaches are being used to tackle new genomes. These include pure WGS methods, the combined approach and variations on that methodology. The future application of the different procedures depends on the target genome sizes, the expected degree of heterogeneity (that is, polymorphism) in the organism to be sequenced, and the preferences of the individual centre. So far, all the genomes that have been analysed by RGSP consortium members have been of high quality and we anticipate that this will continue as the benefits and disadvantages of different approaches are further studied and analysed.

The rat genome data have improved the utility of the rat model enormously. Now that near-complete knowledge of the rat gene content is realizable, individual researchers have a data source for the rat 'parts list' that can be explored with the high degree of confidence and precision that is appropriate for biomedical research. A similar improvement has been made in the resources for physical and genetic mapping, because the relative position of individual markers is now known with high confidence and there are now computational resources to bridge the process of genetic association with gene modelling and experimental investigation. These advances have been reflected by measured increases in the use of all the rat-specific public genome data sets that can be accessed online, as well as by the informally assessed increases in overall 'genomic' research of this model.

The expected benefit of a third mammalian sequence providing an outgroup by which to discriminate the timing of events that had already been noted between mouse and human was fully realized. Using the three sequences and other partial data sets from additional organisms, it was possible to measure some of the overall faster rate of evolutionary change in the rodent lineage shared by mice and rats, as well as the peculiar acceleration of some aspects of rat-specific evolution. The observation of specific expanded gene families in the rat should provide material for targeted studies for some time.

At this time there is no plan to further upgrade or finish the rat genome sequence. This programme decision is a consequence of the high cost of converting draft sequence to finished data, and the pressing need to analyse new genomes. However, as the distant objective of very-low-cost sequencing or other advances that can improve draft sequences inexpensively are realized, it might be envisioned that a rat sequence that approaches the quality of the current human data will be produced. A finished rat genome may answer many questions, as specific clues already show that areas of the genome that are most difficult to resolve in a random sequencing project are also those areas that are most dynamic, and therefore of high potential interest in an evolutionary context.

Despite the advances represented here, we are clearly still at the beginning of the full analysis of the mammalian genome and its complex evolutionary history. Much of the additional data that are required to complete this story will be from other genomes, distantly related to rat. Nevertheless, a considerable body of data remains to be developed from this species. In addition to the distant prospect of a finished rat genome, analysis of other rat strains may yield genome-wide polymorphism data, while targeted efforts to generate cDNA clone collections will provide rat-specific reagents for routine use in research. Together with the ongoing efforts to fully develop methods to genetically manipulate whole rats and provide effective 'gene knockouts', the current and future rat genome resources will ensure a place for this organism in genomic and biomedical research for some time. □

Methods

DNA sequencing and data access

Paired-end reads from BAC and WGS libraries were produced as previously described[2,203]. Unprocessed sequence reads are available from the NCBI Trace Archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/rattus_norvegicus/); raw eBAC assembly data are available from the BCM-HGSC (http://www.hgsc.bcm.tmc.edu/Rat/); and the released Rnor3.1 assembly is available from the BCM-HGSC (ftp://ftp.hgsc.bcm.tmc.edu/pub/analysis/rat/), the NCBI (ftp://ftp.ncbi.nih.gov/genomes/R_norvegicus), and the UCSC (http://genome.ucsc.edu/downloads.html).

Genome assembly

Assembly of the rat genome by the *Atlas* system is described in detail elsewhere[54]. Earlier assemblies (Rnor2.0/2.1) of the initial data set were based on 40 million total reads and 19,000 BAC skims. These assemblies spanned 2.66 Gb and comprised over 900 ultrabactigs with N_{50} of over 5 Mb. They differed only in the removal of short artefactual duplications from Rnor2.0. Rnor3.1 includes another 1,100 BACs, selected to fill gaps in Rnor2.1. Because of the comprehensive coverage of the genome by Rnor2.0/2.1, it was used for the initial predictions of genes and proteins.

BAC fingerprints

An agarose-gel-based fingerprinting methodology[204–207] was employed to generate *Hind*III fingerprints from 199,782 clones in the CHORI-230 BAC library. The contig assembly was subjected to manual review and editing to refine clone order within contigs and to make merges between contigs, using tools provided in the FPC software[208–210]. Fingerprints for 5,250 RPCI-31 PACs[211] and RPCI-32 BACs were subsequently added to allow correlation between the fingerprint map and a developing YAC map of the rat genome. BAC and PAC clones are available through BACPAC Resources at CHORI (bacpacorders@chori.org).

BAC, PAC and YAC maps

Markers generated from BAC and PAC clones were hybridized against YAC[58] (R.D., Pmatch, unpublished software) and radiation hybrid libraries[61,212] to produce independent maps that were subsequently combined. Genetic markers from two rat

genetic maps[61] and the radiation hybrid map[59] were aligned to the Rnor3.1 assembly using BLAT[123] (when sequence was available) or electronic polymerase chain reaction (EPCR)[213].

Finished sequence used for quality assessment of the assembly

To assess the accuracy of the *Atlas* assembly, the Rnor3.1 sequence was compared to 13 Mb of sequences that had been finished to high quality.

Large-scale rearrangements

We compared these assemblies: Human (April 2003, NCBI build 33); Mouse (February 2003, NCBI build 30); and Rat (June 2003, Rnor3.1). Repeats were masked using RepeatMasker (A.S. & P. Green, unpublished work; see http://ftp.genome.washington.edu/RM/RepeatMasker.html) and TandemRepeatFinder[214]. Local alignments were produced using PatternHunter[70] (Supplementary Information). Repeat contamination was removed and the remaining similarities combined into two- and three-way anchors[73] and synteny blocks produced at various resolutions using GRIMM-Synteny[71].

Genome-wide visualization of conserved synteny

Pairwise comparisons of the genomes of human, mouse and rat using MULTIZ[69,215], MLAGAN[216,217], MAVID[110], PatternHunter[70] and Pash[72] were merged into blocks of conserved synteny[69,71,72], and the 1-Mb-resolution images were displayed using the Virtual Genome Painting method (M.L.G.-G. *et al.*, unpublished work; http://www.genboree.org).

Rat segmental duplications

Segmental duplications >5 kb were identified, extracted and aligned as described[218], and paralogous sequence relationships were assessed using PARASIGHT visualization software (J.A.B., unpublished work; Supplementary Information).

Venn diagram

Pairwise and three-way alignments generated using BLASTZ[219] and MULTIZ[215] or HUMOR[215] were analysed to classify each nucleotide in the three genomes by the species with which it aligns: in all three species, aligning between human and rat (but not mouse), between human and mouse (but not rat), or between mouse and rat (but not human). Other nucleotides are species-specific; unassigned nucleotides occupying gaps in the genome assemblies were excluded. On the basis of output from RepeatMasker[164] and RepeatDater[89], nucleotides were assigned to categories (of non-repetitive, repetitive with a certain ancestry, or repetitive but unassigned) and counted. See Supplementary Table SI-1 for details.

Gene prediction

ENSEMBL transcript models were built from 28,478 rodent proteins that were aligned to the genome using a combination of Pmatch (R.D., unpublished software), BLAST[220] and GeneWise[221]. Models based on 5,083 vertebrate proteins were added in regions without rodent-protein-based models. UTRs were added using 11,170 transcripts built from 8,615 different rat cDNAs aligned to the genome using BLAT, with coverage ≥90% and identity ≥95%. This procedure (as described[112] but without GENSCAN predictions), gave rise to 18,241 genes and 20,373 transcripts. This is the protein-based gene set. Rat and mouse cDNA and rat EST-based gene sets were also built. See Supplementary Information for details.

Non-processed pseudogene identification

Human and mouse genes related by 1:1 orthology and lacking an apparent rat orthologue were considered. See Supplementary Information for details.

High-resolution analyses of chromosome 10

These were performed predominantly on the whole genome alignments[217]. Plots in Fig. 9 were generated by sliding windows of width 2 Mb and a step size of 400 kb (total = 277 windows). See Supplementary Information for details.

Received 31 December 2003; accepted 20 February 2004; doi:10.1038/nature02426.

1. Darwin, C. *On The Origin of Species by Means of Natural Selection* 1st edn, Ch. 4, 108 (John Murray, London, 1859).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
4. Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol. Biol. Evol.* **18**, 777–791 (2001).
5. Springer, M. S., Murphy, W. J., Eizirik, E. & O'Brien, S. J. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100**, 1056–1061 (2003).
6. Canby, T. Y. The rat, lapdog of the devil. *Nat. Geogr.* July, 60–87 (1977).
7. Robinson, R. *Genetics of the Norway Rat* (Pergamon, Oxford, 1965).
8. Barnett, S. A. *The Story of Rats. Their Impact on Us, and Our Impact on Them* Ch. 2, 17–18 (Allen and Unwin, Crows Nest, Australia, 2002).
9. Hedrich, H. J. in *History, Strains, and Models in the Laboratory Rat* (ed. Krinke, G. J.) 3–16 (Academic, San Diego, 2000).
10. Lindsey, J. R. in *The Laboratory Rat* (eds Baker, H. J., Lindsey, J. R. & Weisbroth, S. H.) 1–36 (Academic, New York, 1979).
11. Greenhouse, D. D., Festing, M. F. W., Hasan, S. & Cohen, A. L. in *Genetic Monitoring of Inbred Strains of Rats* (ed. Hedrich, H. J.) 410–480 (Gustav Fischer, Stuttgart, 1990).
12. Kuntz, C. *et al.* Comparison of laparoscopic versus conventional technique in colonic and liver resection in a tumor-bearing small animal model. *Surg. Endosc.* **16**, 1175–1181 (2002).
13. Kitagawa, K., Hamada, Y., Nakai, K., Kato, Y. & Okumura, T. Comparison of one- and two-step procedures in a rat model of small bowel transplantation. *Transplant. Proc.* **34**, 1030–1032 (2002).
14. Sauve, Y., Girman, S. V., Wang, S., Keegan, D. J. & Lund, R. D. Preservation of visual responsiveness in the superior colliculus of RCS rats after retinal pigment epithelium cell transplantation. *Neuroscience* **114**, 389–401 (2002).
15. Wang, H. *et al.* Attenuation of acute xenograft rejection by short-term treatment with LF15–0195 and monoclonal antibody against CD45RB in a rat-to-mouse cardiac transplantation model. *Transplantation* **75**, 1475–1481 (2003).
16. Alves, A. *et al.* Total vascular exclusion of the liver enhances the efficacy of retroviral-mediated associated thymidine kinase and interleukin-2 genes transfer against multiple hepatic tumors in rats. *Surgery* **133**, 669–677 (2003).
17. Liu, M. Y., Poellinger, L. & Walker, C. L. Up-regulation of hypoxia-inducible factor 2α in renal cell carcinoma associated with loss of Tsc-2 tumor suppressor gene. *Cancer Res.* **63**, 2675–2680 (2003).
18. Jin, X. *et al.* Effects of leptin on endothelial function with OB-Rb gene transfer in Zucker fatty rats. *Atherosclerosis* **169**, 225–233 (2003).
19. Ravingerova, T., Neckar, J. & Kolar, F. Ischemic tolerance of rat hearts in acute and chronic phases of experimental diabetes. *Mol. Cell. Biochem.* **249**, 167–174 (2003).
20. Taylor, J. R. *et al.* An animal model of Tourette's syndrome. *Am. J. Psychiatry* **159**, 657–660 (2002).
21. Smyth, M. D., Barbaro, N. M. & Baraban, S. C. Effects of antiepileptic drugs on induced epileptiform activity in a rat model of dysplasia. *Epilepsy Res.* **50**, 251–264 (2002).
22. McBride, W. J. & Li, T. K. Animal models of alcoholism: neurobiology of high alcohol-drinking behavior in rodents. *Crit. Rev. Neurobiol.* **12**, 339–369 (1998).
23. Crisci, A. R. & Ferreira, A. L. Low-intensity pulsed ultrasound accelerates the regeneration of the sciatic nerve after neurotomy in rats. *Ultrasound Med. Biol.* **28**, 1335–1341 (2002).
24. Ozkan, O. *et al.* Reinnervation of denervated muscle in a split-nerve transfer model. *Ann. Plast. Surg.* **49**, 532–540 (2002).
25. Fray, M. J., Dickinson, R. P., Huggins, J. P. & Occleston, N. L. A potent, selective inhibitor of matrix metalloproteinase-3 for the topical treatment of chronic dermal ulcers. *J. Med. Chem.* **46**, 3514–3525 (2003).
26. Petratos, P. B. *et al.* Full-thickness human foreskin transplantation onto nude rats as an *in vivo* model of acute human wound healing. *Plast. Reconstr. Surg.* **111**, 1988–1997 (2003).
27. Hussar, P. *et al.* Bone healing models in rat tibia after different injuries. *Ann. Chir. Gynaecol.* **90**, 271–279 (2001).
28. Yang, T. D., Pei, J. S., Yang, S. L., Liu, Z. Q. & Sun, R. L. Medical prevention of space motion sickness–animal model of therapeutic effect of a new medicine on motion sickness. *Adv. Space Res.* **30**, 751–755 (2002).
29. Forte, A. *et al.* Stenosis progression after surgical injury in Milan hypertensive rat carotid arteries. *Cardiovasc. Res.* **60**, 654–663 (2003).
30. Komamura, K. *et al.* Differential gene expression in the rat skeletal and heart muscle in glucocorticoid-induced myopathy: analysis by microarray. *Cardiovasc. Drugs Ther.* **17**, 303–310 (2003).
31. McBride, M. W. *et al.* Functional genomics in rodent models of hypertension. *J. Physiol. (Lond.)* **554**, 56–63 (2004).
32. Kasteleijn-Nolst Trenite, D. G. & Hirsch, E. Levetiracetam: preliminary efficacy in generalized seizures. *Epileptic Disord.* **5**, S39–S44 (2003).
33. Malik, A. S. *et al.* A novel dehydroepiandrosterone analog improves functional recovery in a rat traumatic brain injury model. *J. Neurotrauma* **20**, 463–476 (2003).
34. Kostrubsky, V. E. *et al.* Evaluation of hepatotoxic potential of drugs by inhibition of bile acid transport in cultured primary human hepatocytes and intact rats. *Toxicol. Sci.* **76**, 220–228 (2003).
35. Lindon, J. C. *et al.* Contemporary issues in toxicology: the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.* **187**, 137–146 (2003).
36. Tam, R. C. *et al.* The ribavirin analog ICN 17261 demonstrates reduced toxicity and antiviral effects with retention of both immunomodulatory activity and reduction of hepatitis-induced serum alanine aminotransferase levels. *Antimicrob. Agents Chemother.* **44**, 1276–1283 (2000).
37. Youssef, A. F., Turck, P. & Fort, F. L. Safety and pharmacokinetics of oral lansoprazole in preadolescent rats exposed from weaning through sexual maturity. *Reprod. Toxicol.* **17**, 109–116 (2003).
38. National Institutes of Health. *Network for Large-Scale Sequencing of the Rat Genome* (http://grants2.nih.gov/grants/guide/rfa-files/RFA-HG-00-002.html) (2000).
39. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
40. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
41. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
42. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
43. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
44. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
45. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. & Venter, J. C. On the sequencing and assembly of the human genome. *Proc. Natl Acad. Sci. USA* **99**, 4145–4146 (2002).
46. Waterston, R. H., Lander, E. S. & Sulston, J. E. On the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* **99**, 3712–3716 (2002).
47. Waterston, R. H., Lander, E. S. & Sulston, J. E. More on the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* **100**, 3022–3024 (2003); author reply **(100)**, 3025–3026 (2003).
48. Green, P. Whole-genome disassembly. *Proc. Natl Acad. Sci. USA* **99**, 4143–4144 (2002).
49. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177–189 (2002).

50. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
51. Cheung, J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol.* **4**, R47 [online] (2003).
52. Eichler, E. E. Masquerading repeats: Paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762 (1998).
53. Eichler, E. E. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res.* **11**, 653–656 (2001).
54. Havlak, P. *et al.* The *Atlas* genome assembly system. *Genome Res.* **14**, 721–732 (2004).
55. Osoegawa, K. *et al.* BAC Resources for the rat genome project. *Genome Res.* **14**, 780–785 (2004).
56. Krzywinski, M. *et al.* Integrated and sequence-ordered BAC and YAC-based physical maps for the rat genome. *Genome Res.* **14**, 766–779 (2004).
57. Chen, R., Sodergren, E., Gibbs, R. & Weinstock, G. M. Dynamic building of a BAC clone tiling path for genome sequencing project. *Genome Res.* **14**, 679–684 (2004).
58. Cai, L. *et al.* Construction and characterization of a 10-genome equivalent yeast artificial chromosome library for the laboratory rat, *Rattus norvegicus*. *Genomics* **39**, 385–392 (1997).
59. Kwitek, A. E. *et al.* High density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence. *Genome Res.* **14**, 750–757 (2004).
60. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9**, 1–4 (1999).
61. Steen, R. G. *et al.* A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.* **9** (insert), AP1–AP8 (1999).
62. Misra, S. *et al.* Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* **3**, RESEARCH0083.1-0083.22 [online] (2002).
63. Li, X. & Waterman, M. S. Estimating the repeat structure and length of DNA sequences using L-tuples. *Genome Res.* **13**, 1916–1922 (2003).
64. Riethman, H. *et al.* Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.* **14**, 18–28 (2004).
65. Bayona-Bafaluy, M. P. *et al.* Revisiting the mouse mitochondrial DNA sequence. *Nucleic Acids Res.* **31**, 5349–5355 (2003).
66. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
67. Pevzner, P. & Tesler, G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl Acad. Sci. USA* **100**, 7672–7677 (2003).
68. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
69. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
70. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
71. Pevzner, P. & Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* **13**, 37–45 (2003).
72. Kalafus, K. J., Jackson, A. R. & Milosavljevic, A. Pash: Efficient genome-scale sequence anchoring by positional hashing. *Genome Res.* **14**, 672–678 (2004).
73. Bourque, G., Pevzner, P. A. & Tesler, G. Reconstructiong the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* **14**, 507–516 (2004).
74. Graves, J. A., Gecz, J. & Hameister, H. Evolution of the human X—a smart and sexy chromosome that controls speciation and development. *Cytogenet. Genome Res.* **99**, 141–145 (2002).
75. Bourque, G. & Pevzner, P. A. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* **12**, 26–36 (2002).
76. Murphy, W. J., Bourque, G., Tesler, G., Pevzner, P. & O'Brien, S. J. Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum. Genom.* **1**, 30–40 (2003).
77. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
78. Murphy, W. J., Sun, S., Chen, Z. Q., Pecon-Slattery, J. & O'Brien, S. J. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res.* **9**, 1223–1230 (1999).
79. Ventura, M., Archidiacono, N. & Rocchi, M. Centromere emergence in evolution. *Genome Res.* **11**, 595–599 (2001).
80. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
81. Murphy, W. J., Fronicke, L., O'Brien, S. J. & Stanyon, R. The origin of human chromosome 1 and its homologs in placental mammals. *Genome Res.* **13**, 1880–1888 (2003).
82. Stanyon, R., Stone, G., Garcia, M. & Froenicke, L. Reciprocal chromosome painting shows that squirrels, unlike murid rodents, have a highly conserved genome organization. *Genomics* **82**, 245–249 (2003).
83. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
84. Thomas, J. W. *et al.* Pericentromeric duplications in the laboratory mouse. *Genome Res.* **13**, 55–63 (2003).
85. Horvath, J. E. *et al.* Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol. Biol. Evol.* **20**, 1463–1479 (2003).
86. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13**, 159–172 (2003).
87. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14**, 493–506 (2004).
88. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
89. Yang, S. *et al.* Patterns of insertions and their covariation with substitutions in the rat, mouse and human genomes. *Genome Res.* **14**, 517–527 (2004).
90. Roskin, K. M., Diekhans, M. & Haussler, D. in *Proc. 7th Annu. Int. Conf. Res. Comput. Mol. Biol. (RECOMB 2003)* (eds Vingron, M., Istrail, S., Pevzner, P. & Waterman, M.) doi:10.1145/640075.640109, 257–266 (ACM Press, New York, 2003).
91. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human–mouse genomic alignments. *Cold Spring Harbor Symp. Quant. Biol.* (in the press).
92. Cooper, G. M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
93. Dermitzakis, E. T. *et al.* Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582 (2002).
94. Dermitzakis, E. T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).
95. Nekrutenko, A. Rat–mouse comparisons to identify rodent-specific exons. *Genome Res.* (in the press).
96. Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003).
97. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
98. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
99. Huckins, C. The spermatogonial stem cell population in adult rats. I. Their morphology, proliferation and maturation. *Anat. Rec.* **169**, 533–557 (1971).
100. Clermont, Y. Kinetics of spermatogenesis in mammals: seminiferous epithelium cycle and spermatogonial renewal. *Physiol. Rev.* **52**, 198–236 (1972).
101. Makova, K. D., Yang, S. & Chiaromonte, F. Insertions and deletions are male biased too: A whole-genome analysis in rodents. *Genome Res.* **14**, 567–573 (2004).
102. Sundstrom, H., Webster, M. T. & Ellegren, H. Is the rate of insertion and deletion mutation male biased? Molecular evolutionary analysis of avian and primate sex chromosome sequences. *Genetics* **164**, 259–268 (2003).
103. Chang, B. H. & Li, W. H. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube 1 genes and pseudogenes. *J. Mol. Evol.* **40**, 70–77 (1995).
104. Chang, B. H., Shimmin, L. C., Shyue, S. K., Hewett-Emmett, D. & Li, W. H. Weak male-driven molecular evolution in rodents. *Proc. Natl Acad. Sci. USA* **91**, 827–831 (1994).
105. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
106. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
107. Jensen-Seaman, M. I. *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**, 528–538 (2004).
108. Birdsell, J. A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19**, 1181–1197 (2002).
109. Montoya-Burgos, J. I., Boursot, P. & Galtier, N. Recombination explains isochores in mammalian genomes. *Trends Genet.* **19**, 128–130 (2003).
110. Bray, N. & Pachter, L. MAVID Constrained ancestral alignment of multiple sequence. *Genome Res.* **14**, 693–699 (2004).
111. Yap, V. B. & Pachter, L. Identification of evolutionary hotspots in the rodent genomes. *Genome Res.* **14**, 574–579 (2004).
112. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
113. Vitt, U. *et al.* Identification of candidate disease genes by EST alignments, synteny and expression and verification of Ensembl genes on rat chromosome 1q43-54. *Genome Res.* **14**, 640–650 (2004).
114. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
115. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
116. Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 367–375 (1995).
117. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108–117 (2003).
118. Alexandersson, M., Cawley, S. & Pachter, L. SLAM—Cross-species gene finding with a generalized pair hidden Markov model. *Genome Res.* **13**, 496–502 (2003).
119. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**, 46–54 (2003).
120. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17** (suppl. 1), S140–S148 (2001).
121. Wu, J. Q., Shteynberg, D., Arumugam, M., Gibbs, R. A. & Brent, M. R. Identification of rat genes by TWINSCAN gene prediction, RT–PCR, and direct sequencing. *Genome Res.* **14**, 655–671 (2004).
122. Dewey, C. *et al.* Accurate identification of novel human genes through simultaneous gene prediction in human, mouse and rat. *Genome Res.* **14**, 661–664 (2004).
123. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
124. Puente, X. S. & Lopez-Otin, C. A genomic analysis of rat proteases and protease inhibitors. *Genome Res.* **14**, 609–622 (2004).
125. Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nature Rev. Genet.* **4**, 544–558 (2003).
126. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486–487 (2002).
127. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
128. Wolfe, K. H. & Sharp, P. M. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**, 441–456 (1993).
129. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.* **34**, 177–180 (2003).
130. Nekrutenko, A., Makova, K. D. & Li, W. H. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12**, 198–202 (2002).

131. Nekrutenko, A., Chung, W. Y. & Li, W. H. An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet.* **19,** 306–310 (2003).
132. Taylor, M. S., Ponting, C. P. & Copley, R. R. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* **14,** 555–566 (2004).
133. Green, H. & Wang, N. Codon reiteration and the evolution of proteins. *Proc. Natl Acad. Sci. USA* **91,** 4298–4302 (1994).
134. Levinson, G. & Gutman, G. A. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4,** 203–221 (1987).
135. Alba, M. M. & Guigo, R. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* **14,** 549–554 (2004).
136. Alba, M. M., Santibanez-Koref, M. F. & Hancock, J. M. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol. Biol. Evol.* **16,** 1641–1644 (1999).
137. Burge, C. B., Padgett, R. A. & Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2,** 773–785 (1998).
138. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
139. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12,** 701–709 (2003).
140. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
141. Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nature Rev. Genet.* **3,** 827–837 (2002).
142. Hughes, A. L. *Adaptive Evolution of Genes and Genomes* Ch. 7, 143–179 (Oxford Univ. Press, New York, 1999).
143. Tagle, D. A. *et al.* Embryonic epsilon and γ globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203,** 439–455 (1988).
144. Altschul, S. F. & Lipman, D. J. Protein database searches for multiple alignments. *Proc. Natl Acad. Sci. USA* **87,** 5509–5513 (1990).
145. Gumucio, D. L. *et al.* Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes. *Mol. Cell. Biol.* **12,** 4919–4929 (1992).
146. Hardison, R. *et al.* Comparative analysis of the locus control region of the rabbit β-like gene cluster: HS3 increases transient expression of an embryonic ε-globin gene. *Nucleic Acids Res.* **21,** 1265–1272 (1993).
147. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
148. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13,** 64–72 (2003).
149. Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12,** 832–839 (2002).
150. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2,** 100–109 (2001).
151. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13,** 2507–2518 (2003).
152. Kolbe, D., *et al.* Regulatory potential scores from genome-wide 3-way alignments of human, mouse and rat. *Genome Res.* **14,** 700–707 (2004).
153. Wingender, E. *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29,** 281–283 (2001).
154. Trinklein, N. D., Aldred, S. J., Saldanha, A. J. & Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **13,** 308–312 (2003).
155. Philipsen, S., Pruzina, S. & Grosveld, F. The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the β globin locus control region. *EMBO J.* **12,** 1077–1085 (1993).
156. Reddy, P. M. & Shen, C. K. Protein–DNA interactions *in vivo* of an erythroid-specific, human β-globin locus enhancer. *Proc. Natl Acad. Sci. USA* **88,** 8676–8680 (1991).
157. Strauss, E. C. & Orkin, S. H. *In vivo* protein-DNA interactions at hypersensitive site 3 of the human β-globin locus control region. *Proc. Natl Acad. Sci. USA* **89,** 5809–5813 (1992).
158. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
159. Torrents, D., Suyama, M. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13,** 2559–2567 (2003).
160. Zhang, Z., Harrison, P. & Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12,** 1466–1482 (2002).
161. Mulder, N. J. *et al.* The InterPro Database 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315–318 (2003).
162. Oh, B., Hwang, S. Y., Solter, D. & Knowles, B. B. Spindlin, a major maternal transcript expressed in the mouse during the transition from oocyte to embryo. *Development* **124,** 493–503 (1997).
163. Garcia-Meunier, P., Etienne-Julan, M., Fort, P., Piechaczyk, M. & Bonhomme, F. Concerted evolution in the GAPDH family of retrotransposed pseudogenes. *Mamm. Genome* **4,** 695–703 (1993).
164. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
165. Prak, E. T. & Kazazian, H. H. Jr Mobile elements and the human genome. *Nature Rev. Genet.* **1,** 134–144 (2000).
166. Ostertag, E. M. & Kazazian, H. H. Jr Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35,** 501–538 (2001).
167. Weiner, A. M. SINEs and LINEs: the art of biting the hand that feeds you. *Curr. Opin. Cell Biol.* **14,** 343–350 (2002).
168. Martin, S. L. & Bushman, F. D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **21,** 467–475 (2001).
169. Hayward, B. E., Zavanelli, M. & Furano, A. V. Recombination creates novel L1 (LINE-1) elements in *Rattus norvegicus*. *Genetics* **146,** 641–654 (1997).
170. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35,** 41–48 (2003).
171. Quentin, Y. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Res.* **22,** 2222–2227 (1994).
172. Cantrell, M. A. *et al.* An ancient retrovirus-like element contains hot spots for SINE insertion. *Genetics* **158,** 769–777 (2001).
173. Rothenburg, S., Eiben, M., Koch-Nolte, F. & Haag, F. Independent integration of rodent identifier (ID) elements into orthologous sites of some RT6 alleles of *Rattus norvegicus* and *Rattus rattus*. *J. Mol. Evol.* **55,** 251–259 (2002).
174. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316,** 1033–1040 (2002).
175. Salem, A. H., Kilroy, G. E., Watkins, W. S., Jorde, L. B. & Batzer, M. A. Recently integrated Alu elements and human genomic diversity. *Mol. Biol. Evol.* **20,** 1349–1361 (2003).
176. Salem, A. H. *et al.* Alu elements and hominid phylogenetics. *Proc. Natl Acad. Sci. USA* **100,** 12787–127891 (2003).
177. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21,** 1863–1872 (1993).
178. Benit, L. *et al.* Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J. Virol.* **71,** 5652–5657 (1997).
179. Costas, J. Molecular characterization of the recent intragenomic spread of the murine endogenous retrovirus MuERV-L. *J. Mol. Evol.* **56,** 181–186 (2003).
180. Emes, R. D., Beatson, S. A., Ponting, C. P. & Goodstadt, L. Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Res.* **14,** 591–602 (2004).
181. Young, J. M. *et al.* Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* **11,** 535–546 (2002).
182. Zhang, X. & Firestein, S. The olfactory receptor gene superfamily of the mouse. *Nature Neurosci.* **5,** 124–133 (2002).
183. Rouquier, S., Blancher, A. & Giorgi, D. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl Acad. Sci. USA* **97,** 2870–2874 (2000).
184. Clark, A. J., Hickman, J. & Bishop, J. A 45-kb DNA domain with two divergently orientated genes is the unit of organisation of the murine major urinary protein genes. *EMBO J.* **3,** 2055–2064 (1984).
185. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002).
186. Cavaggioni, A. & Mucignat-Caretta, C. Major urinary proteins, α_{2U}-globulins and aphrodisin. *Biochim. Biophys. Acta* **1482,** 218–228 (2000).
187. Hurst, J. L. *et al.* Individual recognition in mice mediated by major urinary proteins. *Nature* **414,** 631–634 (2001).
188. Danielson, P. B. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Curr. Drug Metab.* **3,** 561–597 (2002).
189. Nelson, D. R. Cytochrome P450 and the individuality of species. *Arch. Biochem. Biophys.* **369,** 1–10 (1999).
190. Scarborough, P. E., Ma, J., Qu, W. & Zeldin, D. C. P450 subfamily CYP2J and their role in the bioactivation of arachidonic acid in extrahepatic tissues. *Drug Metab. Rev.* **31,** 205–234 (1999).
191. Willson, T. M. & Kliewer, S. A. PXR, CAR and drug metabolism. *Nature Rev. Drug Discov.* **1,** 259–266 (2002).
192. Gurates, B. *et al.* WT1 and DAX-1 inhibit aromatase P450 expression in human endometrial and endometriotic stromal cells. *J. Clin. Endocrinol. Metab.* **87,** 4369–4377 (2002).
193. Zhang, Z. *et al.* Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res.* **14,** 580–590 (2004).
194. Lopez-Otin, C. & Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nature Rev. Mol. Cell Biol.* **3,** 509–519 (2002).
195. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).
196. Huang, H. *et al.* Evolutionary conservation of human disease gene orthologs in the rat and mouse genomes. *Genome Biol.* (submitted).
197. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17,** 68–74 (2000).
198. Reddy, P. S. & Housman, D. E. The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.* **9,** 364–372 (1997).
199. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
200. Wade, C. M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420,** 574–578 (2002).
201. Zimdahl, H. *et al.* A SNP map of the rat genome generated from cDNA sequences. *Science* **303,** 807 (2004).
202. Mendell, J. T. & Dietz, H. C. When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* **107,** 411–414 (2001).
203. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
204. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
205. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7,** 1072–1084 (1997).
206. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
207. Schein, J. E. A. in *Bacterial Artificial Chromosomes: Methods and Protocols* (eds Zhao, S. & Stodolsky, M.) 143–156 (Humana, Totowa, New Jersey, 2004).
208. Soderlund, C. I. *et al.* FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13,** 523–535 (1997).
209. Soderlund, C. S. *et al.* Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10,** 1772–1787 (2000).
210. Ness, S. R. *et al.* Assembly of fingerprint contigs: parallelized FPC. *Bioinformatics* **18,** 484–485 (2002).
211. Woon, P. Y. *et al.* Construction and characterization of a 10-fold genome equivalent rat P1-derived artificial chromosome library. *Genomics* **50,** 306–316 (1998).
212. Watanabe, T. K. *et al.* A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genet.* **22,** 27–36 (1999).

213. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* 7, 541–550 (1997).
214. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580 (1999).
215. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715 (2004).
216. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721–731 (2003).
217. Brudno, M. *et al.* Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* 14, 685–692 (2004).
218. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11, 1005–1017 (2001).
219. Schwartz, S. *et al.* MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* 31, 3518–3524 (2003).
220. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402 (1997).
221. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* 10, 547–548 (2000).
222. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324 (1994).
223. Chakrabarti, K. & Pachter, L. Visualization of multiple genome annotations and alignments with the K-BROWSER. *Genome Res.* 14, 716–720 (2004).
224. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425 (1987).
225. Haldi, M. L. *et al.* Construction of a large-insert yeast artificial chromosome library of the rat genome. *Mamm. Genome* 8, 284 (1997).

Supplementary Information accompanies the paper on **www.nature.com/nature.**

Acknowledgements Work at Baylor College of Medicine was supported by a grant from the NHGRI and NHLBI to R.A.G. Work at Genome Therapeutics was supported by grants from the NHGRI to D.S. A.S. acknowledges support from the NIGMS. M.B. acknowledges support from the NIH. N.H. was supported by the NGFN/BMBF (German Ministry for Research and Education). B.J.T. and J.M.Y. are supported by an NIH grant from the NIDCD. K.M.R. and G.M.C. are Howard Hughes Medical Institute Predoctoral Fellows. L.M.D'S., K.M. and K.J.K. are supported by training fellowships from the W. M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology. Work at Case Western Reserve was supported in part by NIH grants to E.E.E. Work at IMIM was supported by a grant from Plan Nacional de I + D (Spain). M.M.A. acknowledges support from programme Ramón y Cajal and a grant from the Spanish Ministry of Science and Technology. Work at Universidad de Oviedo was supported by grants from the European Union, Obra Social Cajastur and Gobierno del Principado de Asturias. Work at Penn State University was supported by NHGRI grants. Work at the University of California Berkeley was supported by a grant from the NIH. Work at the Washington University School of Medicine Genome Sequencing Center and the British Columbia Cancer Agency Genome Sciences Centre was supported by an NIH grant. Work at UCSC and CHORI was supported by the NHGRI.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.G. (agibbs@bcm.tmc.edu). The genomic sequence is available under accession numbers AABR03000000 to AABR03137910 in the international sequence databases (GenBank, DDBJ and EMBL).

Rat Genome Sequencing Project Consortium (Participants are arranged under area of contribution, and then by institution.)

DNA sequencing: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[1], George M. Weinstock (Co-principal Investigator)[1], Michael L. Metzker[1], Donna M. Muzny[1], Erica J. Sodergren[1], Steven Scherer[1], Graham Scott[1], David Steffen[1], Kim C. Worley[1], Paula E. Burch[1], Geoffrey Okwuonu[1], Sandra Hines[1], Lora Lewis[1], Christine DeRamo[1], Oliver Delgado[1], Shannon Dugan-Rocha[1], George Miner[1], Margaret Morgan[1], Alicia Hawes[1], Rachel Gill[1]; **Celera** Robert A. Holt (Principal Investigator)[2,3], Mark D. Adams[3,4], Peter G. Amanatides[3,5], Holly Baden-Tillson[3,6], Mary Barnstead[3,7], Soo Chin[3], Cheryl A. Evans[3], Steve Ferriera[3,8], Carl Fosler[3], Anna Glodek[3,9], Zhiping Gu[3], Don Jennings[3], Cheryl L. Kraft[3,10], Trixie Nguyen[3], Cynthia M. Pfannkoch[3,6], Cynthia Sitter[3,11], Granger G. Sutton[3], J. Craig Venter[3,8], Trevor Woodage[3]; **Genome Therapeutics** Douglas Smith (Principal Investigator)[12,13], Hong-Mei Lee[12], Erik Gustafson[12,13], Patrick Cahill[12], Arnold Kana[12], Lynn Doucette-Stamm[12,13], Keith Weinstock[12], Kim Fechtel[12]; **University of Utah** Robert B. Weiss (Principal Investigator)[14], Diane M. Dunn[14]; **NISC Comparative Sequencing Program, NHGRI** Eric D. Green[15], Robert W. Blakesley[15], Gerard G. Bouffard[15]

BAC library production: Children's Hospital Oakland Research Institute Pieter J. de Jong (Principal Investigator)[16], Kazutoyo Osoegawa[16], Baoli Zhu[16]

BAC fingerprinting: British Columbia Cancer Agency, Canada's Michael Smith Genome Sciences Centre Marco Marra (Principal Investigator)[2], Jacqueline Schein (Principal Investigator)[2], Ian Bosdet[2], Chris Fjell[2], Steven Jones[2], Martin Krzywinski[2], Carrie Mathewson[2], Asim Siddiqui[2], Natasja Wye[2]; **Genome Sequencing Center, Washington University School of Medicine** John McPherson[1,17]

BAC end sequencing: TIGR Shaying Zhao (Principal Investigator)[18], Claire M. Fraser[18], Jyoti Shetty[18], Sofiya Shatsman[18], Keita Geer[18], Yixin Chen[18], Sofyia Abramzon[18], William C. Nierman[18]

Sequence assembly: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[1], George M. Weinstock (Principal Investigator)[1], Paul H. Havlak[1], Rui Chen[1], K. James Durbin[1], Amy Egan[1], Yanru Ren[1], Xing-Zhi Song[1], Bingshan Li[1], Yue Liu[1], Xiang Qin[1]

Analysis and annotation: Affymetrix Simon Cawley[19]; **Baylor College of Medicine** George M. Weinstock (Coordinator)[1], Kim C. Worley (Overall Coordinator)[1], A. J. Cooney[20], Richard A. Gibbs[1], Lisa M. D'Souza[1], Kirt Martin[1], Jia Qian Wu[1], Manuel L. Gonzalez-Garay[1], Andrew R. Jackson[1], Kenneth J. Kalafus[1,58], Michael P. McLeod[1], Aleksandar Milosavljevic[1], Davinder Virk[1], Andrei Volkov[1], David A. Wheeler[1], Zhengdong Zhang[1]; **Case Western Reserve University** Jeffrey A. Bailey[4], Evan E. Eichler[4], Eray Tuzun[4]; **EBI, Wellcome Trust Genome Campus** Ewan Birney[21], Emmanuel Mongin[21], Abel Ureta-Vidal[21], Cara Woodwark[21]; **EMBL, Heidelberg** Evgeny Zdobnov[22], Peer Bork[22,23], Mikita Suyama[22], David Torrents[22]; **Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg** Marina Alexandersson[24]; **Fred Hutchinson Cancer Research Center** Barbara J. Trask[25], Janet M. Young[25]; **Genome Therapeutics** Douglas Smith (Principal Investigator)[12,13], Hui Huang[12], Kim Fechtel[12], Huajun Wang[12], Heming Xing[12], Keith Weinstock[12]; **Incyte Corporation** Sue Daniels[26], Darryl Gietzen[26], Jeanette Schmidt[26], Kristian Stevens[26], Ursula Vitt[26], Jim Wingrove[26]; **Institut Municipal d'Investigacio Medica, Barcelona** Francisco Camara[27], M. Mar Albà[27], Josep F. Abril[27], Roderic Guigo[27]; **The Institute for Systems Biology** Arian Smit[28]; **Lawrence Berkeley National Laboratory** Inna Dubchak[29,30], Edward M. Rubin[29,30], Olivier Couronne[29,30], Alexander Poliakov[29]; **Max Delbrück Center for Molecular Medicine** Norbert Hübner[23], Detlev Ganten[23], Claudia Goesele[23,31], Oliver Hummel[23,31], Thomas Kreitler[23,31], Young-Ae Lee[23], Jan Monti[23], Herbert Schulz[23], Heike Zimdahl[23];

Max Planck Institute for Molecular Genetics, Berlin Heinz Himmelbauer[31], Hans Lehrach[31]; **Medical College of Wisconsin** Howard J. Jacob (Principal Investigator)[32], Susan Bromberg[33], Jo Gullings-Handley[32], Michael I. Jensen-Seaman[32], Anne E. Kwitek[32], Jozef Lazar[32], Dean Pasko[33], Peter J. Tonellato[32], Simon Twigger[32]; **MRC Functional Genetics Unit, University of Oxford** Chris P. Ponting (Leader, Genes and Proteins Analysis Group)[34], Jose M. Duarte[34], Stephen Rice[34], Leo Goodstadt[34], Scott A. Beatson[34], Richard D. Emes[34], Eitan E. Winter[34], Caleb Webber[34]; **MWG-Biotech** Petra Brandt[35], Gerald Nyakatura[35]; **Pennsylvania State University** Margaret Adetobi[36], Francesca Chiaromonte[36], Laura Elnitski[36], Pallavi Eswara[36], Ross C. Hardison[36], Minmei Hou[36], Diana Kolbe[36], Kateryna Makova[36], Webb Miller[36], Anton Nekrutenko[36], Cathy Riemer[36], Scott Schwartz[36], James Taylor[36], Shan Yang[36], Yi Zhang[36]; **Roche Genetics and Roche Center for Medical Genomics** Klaus Lindpaintner[37]; **Sanger Institute** T. Dan Andrews[38], Mario Caccamo[38], Michele Clamp[38], Laura Clarke[38], Valerie Curwen[38], Richard Durbin[38], Eduardo Eyras[38], Stephen M. Searle[38]; **Stanford University** Gregory M. Cooper (Co-Leader, Evolutionary Analysis Group)[39], Serafim Batzoglou[40], Michael Brudno[40], Arend Sidow[39], Eric A. Stone[39]; **The Center for the Advancement of Genomics** J. Craig Venter[3,8]; **University of Arizona** Bret A. Payseur[41]; **Université de Montréal** Guillaume Bourque[42]; **Universidad de Oviedo** Carlos López-Otín[43], Xose S. Puente[43]; **University of California, Berkeley** Kushal Chakrabarti[44], Sourav Chatterji[44], Colin Dewey[44], Lior Pachter[45], Nicolas Bray[45], Von Bing Yap[45], Anat Caspi[46]; **University of California, San Diego** Glenn Tesler[47], Pavel A. Pevzner[48]; **University of California, Santa Cruz** David Haussler (Co-Leader, Evolutionary Analysis Group)[49], Krishna M. Roskin[50], Robert Baertsch[50], Hiram Clawson[50], Terrence S. Furey[50], Angie S. Hinrichs[50], Donna Karolchik[50], William J. Kent[50], Kate R. Rosenbloom[50], Heather Trumbower[50], Matt Weirauch[36,50]; **University of Wales College of Medicine** David N. Cooper[51], Peter D. Stenson[51]; **University of Western Ontario** Bin Ma[52]; **Washington University** Michael Brent[53], Manimozhiyan Arumugam[53], David Shteynberg[53]; **Wellcome Trust Centre for Human Genetics, University of Oxford** Richard R. Copley[54], Martin S. Taylor[54]; **The Wistar Institute** Harold Riethman[55], Uma Mudunuri[55]

Scientific management: Jane Peterson[56], Mark Guyer[56], Adam Felsenfeld[56], Susan Old[57], Stephen Mockrin[57] & Francis Collins[56]

Affiliations for participants: 1, Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, MS BCM226, One Baylor Plaza, Houston, Texas 77030, USA (http://www.hgsc.bcm.tmc.edu); 2, British Columbia Cancer Agency, Canada's Michael Smith Genome Sciences Centre, 600 W 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada (http://www.bcgsc.ca); 3, Celera, 45 West Gude Drive, Rockville, Maryland 20850, USA; 4, Department of Genetics and the Center for Computational Genomics, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA; 5, DSM Pharmaceuticals Inc., 5900 NW Greenville Blvd, Greenville, North Carolina 27834, USA; 6, The Institute for Biological Energy Alternatives (IBEA), 1901 Research Blvd, Rockville, Maryland 20850, USA; 7, Intronn, Inc., 910 Clopper Road, South Building, Gaithersburg, Maryland 20878, USA; 8, The Center for the Advancement of Genomics (TCAG), 1901 Research Blvd, Suite 600, Rockville, Maryland 20850, USA; 9, Avalon Pharmaceuticals, 20358 SenecaMeadows Parkway, Germantown, Maryland 20876, USA; 10, Basic Immunology Branch, Division of Allergy, Immunology and Transplantation, National Institute of Allergy and Infectious Diseases (NIAID), NIH, DHHS, 6610 Rockledge Blvd, Room 3005, Bethesda, Maryland 20892-7612, USA; 11, DynPort Vaccine Company, LLC, 64 Thomas Jefferson Drive, Frederick, Maryland 21702, USA; 12, Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453, USA; 13, Agencourt Bioscience Corporation, 100 Cummings Center, Beverly, Massachusetts 01915, USA; 14, Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA; 15, NIH Intramural Sequencing Center (NISC) and Genome Technology Branch, National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, Maryland 20892, USA; 16, BACPAC Resources, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA (http://bacpac.chori.org); 17, Genome Sequencing Centre, Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, Missouri 63108, USA (http://genome.wustl.edu); 18, The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, Maryland 20850, USA (http://www.tigr.org); 19, Affymetrix, 6550 Vallejo St, Suite 100, Emeryville, California 94608, USA; 20, Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA; 21, EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK; 22, EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany; 23, Max Delbrück Center for Molecular Medicine (MDC), Experimental Genetics of Cardiovascular Disease, Robert-Rössle-Strasse 10, Berlin 13125, Germany (http://www.mdc-berlin.de/ratgenome/); 24, Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Chalmers Science Park, S-412 88 Gothenburg, Sweden; 25, Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., C3-168, Seattle, Washington 98109, USA (http://www.fhcrc.org/labs/trask/); 26, Incyte Corporation, 3160 Porter Drive, Palo Alto, California 94304, USA (http://www.incyte.com); 27, Grup de Recerca en Informàtica Biomèdia, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, and Programa de Bioinfomatica i Genomica, Centre de Regulacio Genomica, C/ Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain; 28, Computational Biology Group, The Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA; 29, Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, California 94720, USA (http://www.lbl.gov); 30, US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA (http://jgi.doe.gov); 31, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, Berlin 14195, Germany; 32, Human and Molecular Genetics Center, Bioinformatics Research Center, and Department of Physiology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA; 33, Rat Genome Database, Bioinformatics Research Center, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA; 34, MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK; 35, MWG-Biotech, Anzinger Strasse 7a, Ebersberg 85560, Germany; 36, Center for Comparative Genomics and Bioinformatics, Huck Institutes of Life Sciences, Departments of Biology, Statistics, Biochemistry and Molecular Biology, Computer Science and Engineering, and Health Evaluation Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 37, Roche Genetics and Roche Center for Medical Genomics, F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland; 38, Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 39, Departments of Pathology and Genetics, Stanford University, Stanford, California 94305, USA; 40, S256 James H. Clark Center, Department of Computer Science, Stanford University, Stanford, California 94305, USA; 41, Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA; 42, Centre de Recherches Mathématiques, Université de Montréal, 2920 Chemin de la tour, Montréal, Quebec H3T 1J8, Canada (http://www.crm.umontreal.ca); 43, Departamento de Bioquimica y Biologia Molecular, Instituto Universitario de Oncologia, Universidad de Oviedo, 33006 Oviedo, Spain (http://web.uniovi.es/degradome); 44, Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, California 94720, USA; 45, Department of Mathematics, University of California Berkeley, Berkeley, California 94720, USA; 46, Bioengineering Graduate Group, University of California Berkeley, Berkeley, California 94720, USA; 47, University of California, San Diego, Department of Mathematics, 9500 Gilman Drive, San Diego, California 92093-0112, USA (http://www-cse.ucsd.edu/groups/bioinformatics); 48, University of California, San Diego, Department of Computer Science and Engineering, 9500 Gilman Drive, San Diego, California 92093-0114, USA (http://www-cse.ucsd.edu/groups/bioinformatics); 49, Howard Hughes Medical Institute, Center for Biomolecular Science & Engineering, Mailstop SOE, Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA; 50, UCSC Genome Bioinformatics Group, Center for Biomolecular Science and Engineering, Mailstop SOE, Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA; 51, Institute of Medical Genetics, University of Wales College of Medicine, Heath Park, Cardiff, CF14 4XN, UK; 52, Department of Computer Science, University of Western Ontario, London, Ontario N6A 5B7, Canada; 53, Laboratory for Computational Genomics, Campus Box 1045, Washington University, St Louis, Missouri 63130, USA (http://genes.cse.wustl.edu); 54, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK; 55, The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsyvania 19104, USA; 56, US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA; 57, US National Institutes of Health, National Heart, Lung, and Blood Institute, Bethesda, Maryland 20892, USA; 58, Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

EXHIBIT 5

Motor Neuron Degeneration in Mice That Express a Human Cu,Zn Superoxide Dismutase Mutation

Mark E. Gurney,* Haifeng Pu, Arlene Y. Chiu, Mauro C. Dal Canto, Cynthia Y. Polchow, Denise D. Alexander, Jan Caliendo, Afif Hentati, Young W. Kwon, Han-Xiang Deng, Wenje Chen, Ping Zhai, Robert L. Sufit, Teepu Siddique

Mutations of human Cu,Zn superoxide dismutase (SOD) are found in about 20 percent of patients with familial amyotrophic lateral sclerosis (ALS). Expression of high levels of human SOD containing a substitution of glycine to alanine at position 93—a change that has little effect on enzyme activity—caused motor neuron disease in transgenic mice. The mice became paralyzed in one or more limbs as a result of motor neuron loss from the spinal cord and died by 5 to 6 months of age. The results show that dominant, gain-of-function mutations in SOD contribute to the pathogenesis of familial ALS.

Amyotrophic lateral sclerosis occurs in both sporadic and familial forms and results from the degeneration of motor neurons in the cortex, brainstem, and spinal cord. The disease typically begins in adults as an asymmetric weakness in two or more limbs and then progresses to complete paralysis (*1*). Familial ALS is inherited as an autosomal dominant trait (*2*). About 10% of ALS cases are familial and, of these, ~20% have mutations in Cu,Zn superoxide dismutase (SOD) (*3–5*). SOD catalyzes the dismutation of superoxide radical ($O_2^{\cdot -}$) into hydrogen peroxide and molecular oxygen. Familial ALS patients heterozygous for SOD mutations have 50 to 60% of the normal level of SOD activity in their red blood cells and brains (*4*, *6*).

To explore how mutations in SOD might selectively cause motor neuron degeneration, we produced transgenic mice that express wild-type or mutant forms of human SOD (*7*, *8*). Two mutations were analyzed: an $Ala^4 \rightarrow$ Val substitution (A4V) and a $Gly^{93} \rightarrow$ Ala substitution (G93A) (*3*, *4*). Previously described mice that express wild-type human SOD (NSOD) show no signs of overt motor neuron disease but do have mild pathologic changes in the innervation of muscle that are suggestive of premature aging (*8*, *9*).

Transgenic founder mice were produced by DNX (Princeton, New Jersey) or through the National Transgenic Development Facility (National Institutes of Health). Fertilized eggs for injection were obtained from crosses of (C57BL6 × SJL) F_1 hybrid mice. Founder mice were bred with C57BL6 mice, and their progeny were used for subsequent analysis (*10*). Transgenic mice were identified by polymerase chain reaction amplification of tail DNA (*11*) and were screened for expression of human SOD in red blood cells by an antigen capture enzyme immunoassay (EIA) that used a polyclonal antibody to human SOD and the mouse monoclonal antibody SD-G6 (*12*). The EIA detected human SOD in G93A and NSOD mice, but not in A4V mice. However, Northern (RNA) analysis (*13*) and immunoblots (*14*) developed with a different mouse monoclonal antibody (CZSODF2) demonstrated expression of human *SOD1* mRNA and protein in the brains of G93A, NSOD, and A4V mice (Fig. 1, A to C). Thus, the A4V mutation altered an epitope needed for recognition in the EIA.

The mutations of SOD found in familial ALS alter the stability of human SOD as shown by DNA transfection of cultured cells (*15*). Consistent with those results, we found that the mutant transgenic lines expressed only one-half as much human SOD as did NSOD mice expressing comparable amounts of mRNA (Table 1). In addition, we found that the G93A mutation had little discernible effect on human SOD activity, whereas the A4V mutation greatly reduced enzymatic activity (*15*, *16*). Although we detected enzymatically active mouse-human dimers in NSOD and G93A transgenic mice on SOD activity gels (*17*), we did not detect any active mouse-human A4V dimers. These results are compatible with the finding that recombinant human SOD bearing an $Ala^4 \rightarrow$ Gln substitution is enzymatically inactive (*18*).

Mice from one of the G93A transgenic lines (G1) (Table 1) that expressed the largest amounts of mutant SOD in the brain

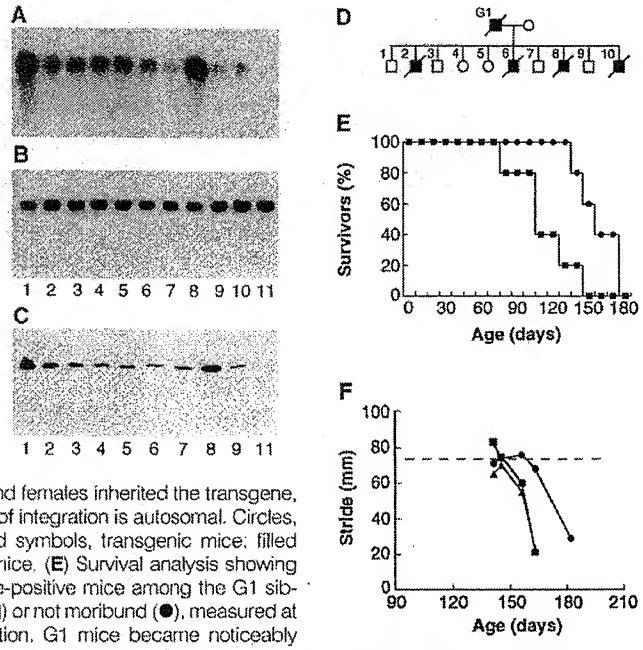


Fig. 1. (**A**) Northern analysis of human *SOD1* mRNA expression in transgenic mouse brain. (**B**) The same membrane hybridized with a probe for G3PDH. (**C**) Expression of human SOD in transgenic brain by immunoblotting. Lanes contain samples from the following mice: 1, G1; 2, G5; 3, G12.199; 4, G20; 5, A1073; 6, A1074; 7, N1026; 8, N1029; 9, N1030; 10, G12.15; and 11, nontransgenic littermate. (**D**) Partial pedigree of the G1 transgenic line. In the F_2 generation, both males and females inherited the transgene, which indicates that the site of integration is autosomal. Circles, female; squares, male; filled symbols, transgenic mice; filled symbols with bar, affected mice. (**E**) Survival analysis showing the percentage of transgene-positive mice among the G1 siblings that are not impaired (■) or not moribund (●), measured at 10-day intervals of observation. G1 mice became noticeably impaired by 121 ± 23 days of age (mean ± SD, n = 5) and moribund by 169 ± 16 days. (**F**) The condition of G1 transgenic mice deteriorated rapidly over the 2-week period before their death, as shown by the shortening of their stride (■, G1.2; ●, G1.6; ▲, G1.8; and dotted line, average stride of normal male mice).

M. E. Gurney, H. Pu, D. D. Alexander, Y. W. Kwon, P. Zhai, Department of Cell and Molecular Biology and Northwestern University Institute of Neuroscience, Northwestern University Medical School, 303 East Chicago Avenue, Chicago, IL 60611, USA.
A. Y. Chiu, Division of Neurosciences, Beckman Research Institute of the City of Hope Medical Center, 1450 East Duarte Road, Duarte, CA 91010, USA.
M. C. Dal Canto, Department of Pathology, Northwestern University Medical School, Chicago, IL 60611, USA.
C. Y. Polchow, Department of Physiology, Northwestern University Medical School, Chicago, IL 60611, USA.
J. Caliendo, A. Hentati, H.-X. Deng, W. Chen, R. L. Sufit, T. Siddique, Department of Neurology and Northwestern University Institute of Neuroscience, Northwestern University Medical School, Chicago, IL 60611, USA.
*To whom correspondence should be addressed.

developed a stereotyped syndrome suggestive of motor neuron disease. The disease has not been observed in any line of NSOD mice expressing wild-type human SOD, nor have symptoms developed in any A4V mouse at comparable ages. At 3 to 4 months of age, G1 mice began to show signs of hind limb weakness (Fig. 1E). They extended their hind legs less than normal when lifted by the base of the tail, their coats developed a coarse appearance suggestive of impaired grooming, and they appeared thin along their flanks. Normal mice have a fairly constant stride of 74 ± 1.6 mm (95% confidence interval, n = 50 mice) when using an alternating gait (*19*). G1 mice had a normal stride at 3 to 4 months of age, but by 5 months of age it deteriorated rapidly (Fig. 1F). Over a span of 2 weeks, the mice became paralyzed in one or more limbs. The founder mouse and four of five transgenic F_1 progeny developed paralysis of one or more hind limbs. A fifth transgenic F_1 mouse (G1.6) retained use of his hind limbs but developed complete paralysis of his right forelimb. The six nontransgenic littermates of these mice showed no signs of disease. All affected mice developed a tremor of the hind limbs when suspended in the air. They had a normal posture when quiet with the hind limbs held in flexion, but after initiating movement, their hind limbs and toes frequently locked in a hyperextended position. Affected mice became moribund by 5 months of age and were killed when they were no longer able to forage for food or water.

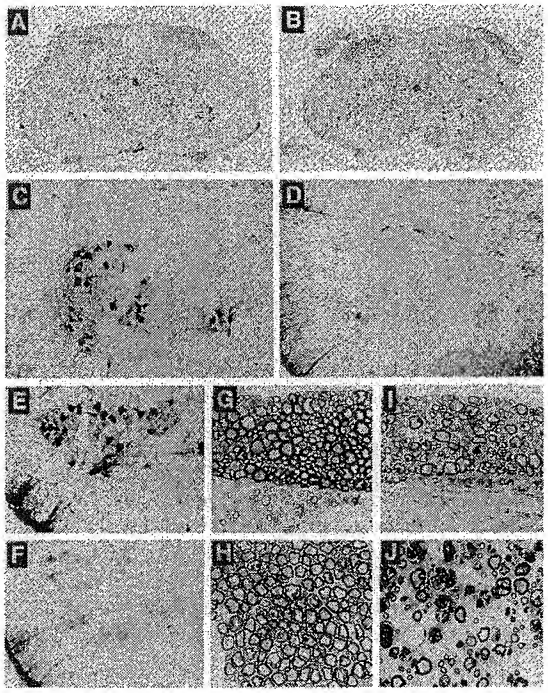


Fig. 2. Loss of spinal motor neurons in affected G1 transgenic mice. Spinal cords from a normal littermate (**A**) and a G1 transgenic mouse (**B**) show loss in the latter of lateral motor columns in the L4 spinal segment (cresylecht-violet stain). (**C** and **D**) Spinal cords from a normal littermate (C) and a G1 transgenic mouse (D), showing loss in the latter of ChAT-positive ventral horn motor neurons in the L3 spinal segment (*27*). Lumbar spinal cords from an N1026 mouse (**E**) and a normal littermate (**F**) show staining of ventral horn motor neurons with an antibody to human SOD (CZ-SODF2). (**G** through **J**) Normal littermate dorsal (G) and ventral (H) lumbar spinal roots and G1 transgenic dorsal (I) and ventral (J) lumbar roots (stained with toluidine blue). The dorsal sensory roots were relatively spared (I), whereas severe loss of myelinated axons, myelin debris, and infiltrating phagocytic cells were apparent in the ventral motor roots (J).

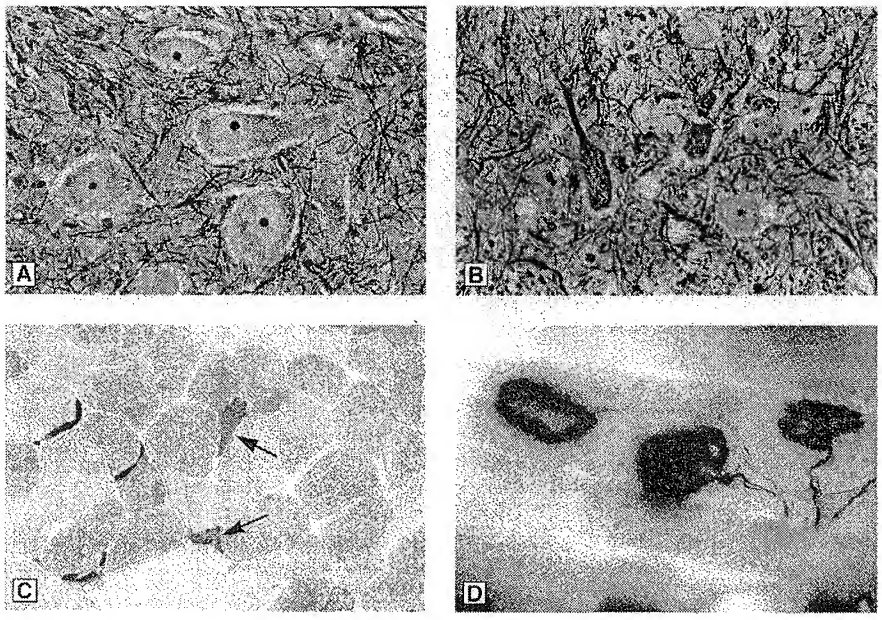


Fig. 3. Pathology in spinal cord and muscle of transgenic G1 mice. (**A** and **B**) Lumbar spinal segments from a normal littermate (A) and a G1 transgenic mouse (B), stained by the Bielschowski technique to reveal neurofibrils. (**C**) Nonspecific esterase stain of gastrocnemius, showing the low frequency of denervated, angulated muscle fibers (arrows) in G1 mice. (**D**) Sprouting and reinnervation of three denervated endplates in the gluteus muscle of a G1 mouse, revealed by a combined silver and cholinesterase stain (*28*).

The founder of the G1 line, all of his transgenic F_1 progeny, and at least one male F_2 mouse developed the same stereotyped syndrome suggestive of motor neuron disease affecting both upper and lower motor neurons. The other lines of G93A transgenic mice (Table 1) expressed smaller amounts of the mutant protein and so far have had normal motor behavior. In G1 mice as well as in humans with ALS (*2*), the onset of the disease is dependent on age, so it is conceivable that the other lines of G93A mice may develop the disease at a later age. However, because the disease is expressed in only one line of mice, we cannot exclude the possibility that the site of integration of the transgene caused the disease syndrome in these mice. Disease is not due simply to overexpression of SOD in the brains of G1 mice, because NSOD mice that express comparable or greater amounts of total brain SOD do not develop the disease (*10*) (Table 1).

Pathological analysis of G1 mice demonstrated a severe loss of choline acetyltransferase (ChAT)–containing spinal motor neurons (Fig. 2, A to D). A few motor neurons appeared normal, but most of the remaining neurons were filled with a neurofibrillar material (Fig. 3) that appeared to be phosphorylated neurofilaments (*20*). The most pronounced changes were observed in the ventral spinal cord, whereas the dorsal spinal cord, especially the substantia gelatinosa, was better preserved. Immunohistochemical staining revealed large amounts of human SOD in ventral horn motor neurons, best shown in NSOD mice (Fig. 2, E and F). In G1 mice, there was severe loss of large, myelinated axons from the ventral motor roots (Fig. 2, H and

Table 1. Expression in brain of human *SOD1* mRNA, human SOD protein, and total SOD enzymatic activity in different transgenic mouse lines. All values are the mean ± SEM (n = 3), except where indicated.

Line	Mutation	Gene copy number*	*SOD1* mRNA (ng)/10 μg of total RNA	Human SOD (ng)/total protein (μg)†	SOD (U)/total protein (μg)
G1	$Gly^{93} \rightarrow Ala$	18.0 ± 2.6	2.5 ± 0.5	4.1 ± 0.54	42.6 ± 2.1
G5		4.0 ± 0.6	0.8 ± 0.1	1.3 ± 0.21	27.0 ± 2.9
G12		2.2 ± 0.8	0.8 ± 0.1	1.1 ± 0.22	19.5 ± 0.8
G20		1.7 ± 0.6	0.8 ± 0.1	0.7 ± 0.06	16.9 ± 0.4
A1073	$Ala^4 \rightarrow Val$	4.7 ± 0.4	1.1 ± 0.1	1.0 ± 0.21‡	14.6 ± 0.4
A1074		3.2 ± 0.2	0.7 ± 0.1	0.9 ± 0.21‡	9.1 ± 0.4
N1029	Wild-type	7.2 ± 2.4	1.5 ± 0.1	6.7 ± 0.76	37.3 ± 1.9
N1026		3.3 ± 1.0	0.4 ± 0.1	0.9 ± 0.11	18.6 ± 0.9
N1030		1.7 ± 0.7	0.3 ± 0.1	0.6 ± 0.16	11.8 ± 0.3
Nontransgenic		–	–	–	10.4 ± 0.5

*Per diploid genome. †The amount of human SOD was determined by EIA. ‡Determined by immunoblotting (mean ± SEM of regression).

J). The dorsal sensory roots appeared relatively spared when compared to the ventral roots; however, scattered swollen axons with dense axoplasm and occasional myelin-laden macrophages were observed at all levels of the spinal cord (Fig. 2, G and I). These changes extended into the central component of the afferent sensory fibers within the dorsal columns of the spinal cord, a pathology also seen in familial ALS (*21*). Severe loss of myelinated axons occurred in intramuscular nerves, but less than 10% of muscle fibers had the characteristics of denervated fibers—that is, a small, angular profile and an esterase-positive phenotype (Fig. 3C).

To investigate whether sprouting and reinnervation compensated for the destruction of motor units caused by the disease, we examined whole mounts of the gluteus muscle of a G1 mouse (Fig. 3D). The muscles showed severe loss of myelinated axons from the intramuscular nerves and consequent reinnervation of muscle fibers by primarily nodal sprouts. Ongoing reinnervation and remodeling of muscle innervation were indicated by the frequency of multiply innervated endplates and by the scarcity of denervated endplates. In one gluteus muscle, two surviving axons in the inferior gluteal nerve appeared sufficient to innervate more than 90% of the myofibers in the muscle. These data suggest that sprouting probably compensates for the loss of motor neurons until late in the course of the disease.

Toxicity by a free-radical mechanism is one plausible explanation for motor neuron death in the G1 transgenic mice and, by implication, in humans with familial ALS. This mechanism could involve the formation of the strong oxidant peroxynitrite ($ONOO^-$) from $O_2^{\cdot -}$ and nitric oxide ($NO\cdot$) free radicals (*22*, *23*). The formation of peroxynitrite and its decomposition into toxic chemical species have been linked to neurotoxicity in cell culture (*24*) and in brain ischemia (*25*). SOD mutations may facilitate this pathway of oxidative damage (*26*). Because formation of peroxynitrite is a second-order reaction that depends on the concentration of $O_2^{\cdot -}$ and $NO\cdot$, decreased SOD activity in familial ALS may also contribute to pathogenesis if the amount of $O_2^{\cdot -}$ in tissues is increased (*4*). Our results indicate that dominant, gain-of-function mutations in SOD play a key role in the pathogenesis of familial ALS.

REFERENCES AND NOTES

1. D. W. Mulder, in *Human Motor Neuron Diseases*, L. P. Rowland, Ed. (Raven, New York, 1982), pp. 15–22.
2. T. Siddique, *Adv. Neurol.* **56**, 227 (1991).
3. D. R. Rosen *et al.*, *Nature* **362**, 59 (1993).
4. H.-X. Deng *et al.*, *Science* **261**, 1047 (1993).
5. T. Siddique, unpublished observations.
6. A. C. Bowling, J. B. Schulz, R. H. Brown Jr., M. F. Beal, *J. Neurochem.* **61**, 2322 (1993).
7. The A4V mutation was introduced into exon 1 of the human *SOD1* gene by two-primer mutagenesis with the polymerase chain reaction (PCR); the template for mutagenesis was a Sty I–Stu I fragment encompassing exon 1. The G93A mutation was cloned in a Hind III and Nsi I fragment encompassing exon 4 that was amplified from the genomic DNA of family 3-192 (*3*). These fragments were used to reassemble a complete 14.5-kb Eco RI–Bam HI fragment of the *SOD1* gene [R. A. Hallewell, J. P. Puma, G. T. Mullenbach, R. C. Najarian, in *Superoxide and Superoxide Dismutase in Chemistry, Biology and Medicine*, G. Rotilo, Ed. (Elsevier, New York, 1986), pp. 249–256] in two more steps. Exons 1 and 4 of the transgenes were sequenced to verify that they contained only the desired mutation. The 14.5-kb Eco RI–Bam HI *SOD1* transgene directs tissue-specific expression of human SOD in mice under control of the endogenous human promoter (*8*).
8. C. J. Epstein *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8044 (1987).
9. I. Ceballos-Picot *et al.*, *Brain Res.* 552, 198 (1991); K. B. Avraham *et al.*, *Cell* **54**, 823 (1988); K. B. Avraham, H. Sugarman, S. Rotshenker, Y. Groner, *J. Neurocytol.* **20**, 208 (1991).
10. Mice were housed in microisolator cages within a barrier facility. Frequent monitoring revealed no evidence for infection by viral or bacterial pathogens.
11. The primers described (*3*) were used for identification of transgenic mice by PCR. Transgene copy number was estimated by Southern (DNA) DNA hybridization. Denatured DNA (10 μg) isolated from mouse tails or human placenta was transferred to a nitrocellulose membrane together with a dilution series of the cloned *SOD1* gene. The membrane was hybridized with a random-primed, ^{32}P-labeled probe to sequences within the 3′ untranslated region of the 0.9-kb human *SOD1* complementary DNA (cDNA); these sequences are specific to the human transgene. Bound radioactivity was quantitated by phosphor image analysis, and linear regression was used to calculate transgene copy number.
12. The EIA was constructed with a goat immunoglobulin G (IgG) antibody to human SOD (Chiron, Emeryville, CA) and a mouse monoclonal antibody designated SD-G6 (Sigma, St. Louis, MO). Recombinant human SOD (Chiron) was used as a standard. Samples were diluted to within the log-linear range of the assay (0.1 to 1.5 ng of human SOD per well). There was no cross-reactivity with mouse SOD.
13. Northern RNA hybridization was performed with 10 μg of total brain RNA. The membrane was hybridized with a random-primed, ^{32}P-labeled probe specific for the 3′ untranslated region of the human *SOD1* cDNA. Quantitation standards (a 0.9-kb sense human *SOD1* cDNA) were loaded on the gel with 10 μg of yeast RNA as a carrier, and the hybridization signal was analyzed by phosphor image analysis. To control for RNA loading variations, we rehybridized the blot with a glyceraldehyde 3-phosphate dehydrogenase (G3PDH) cDNA probe.
14. Samples containing 2 μg of soluble brain protein were subjected to electrophoresis through 10% SDS-polyacrylamide gels, transferred to a nitrocellulose membrane, and probed with antibody CZSODF2. Bound antibody was detected with a biotinylated horse antibody to mouse IgG and a Vector ABC kit. The membrane was developed with an enhanced chemiluminescence kit (Amersham, Arlington Heights, IL), and the chemiluminescence was quantitated by film densitometry. The amount of human SOD in brain extracts of A4V transgenic mice was determined by comparison to recombinant SOD standards in adjacent lanes.
15. D. R. Borchelt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
16. Mouse brains were homogenized in cold 10 mM tris HCl (pH 7.5) and 10 mM β-mercaptoethanol. After centrifugation at 50,000*g* for 15 min at 4°C, the protein content of the supernatant was measured by a bicinchoninic acid assay (Pierce, Rockford, IL). We assayed total SOD activity within brain extracts in microwells by measuring the inhibition of nitroblue tetrazolium reduction [D. R. Spitz and L. W. Oberley, *Anal. Biochem.* **179**, 8 (1989)]. Wells were monitored kinetically, and a $(V_{max}/V) - 1$ transform (where V is velocity) [K. Asada, M. Takahashi, M. Nagate, *Agric. Biol. Chem.* 38, 471 (1974)] was used to linearize the data. Recombinant human SOD had an activity of 6 U per nanogram. The contribution of Mn SOD in the sample was determined in the presence of 5 mM sodium cyanide and was ~2% of the total SOD activity in the brain extract.
17. O. Elroy-Stein, Y. Bernstein, Y. Groner, *EMBO J.* 5, 615 (1986).
18. R. A. Hallewell *et al.*, *Nucleic Acids Res.* **13**, 2017 (1985).
19. Mice were trained to walk up a 75-cm, U-shaped ramp that was inclined at one end against the wire lid of their cage. Testing was performed in a horizontal, laminar flow hood to maintain barrier conditions. A bright lamp was placed at the base of the ramp, and the cage lid was left in semidarkness. The ramp obscured each mouse's view of the laminar flow hood and surrounding room. Testing was initiated by allowing the mice 1 to 2 min to explore the cage lid and the top of the

ramp. The hind feet of the mice were painted with children's poster paints of contrasting colors. The tracks left by the mice as they ran up the ramp were recorded on paper tape.
20. Degenerating neurons were positive for immunohistochemical staining with SMI-31 monoclonal antibody (Sternberger Monoclonal Antibodies, Baltimore, MD) to phosphorylated neurofilaments, although the small number of motor neurons remaining in affected spinal cords and their marked pathology require confirmation of this result.
21. W. K. Engel, L. T. Kurland, I. Klatzo, *Brain* **82**, 203 (1959); A. Hirano, L. T. Kurland, G. P. Sayre, *Arch. Neurol.* **16**, 232 (1967).
22. J. S. Beckman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1624 (1990).
23. H. Ischiropoulos *et al.*, *Arch. Biochem. Biophys.* **298**, 431 (1992).
24. S. A. Lipton *et al.*, *Nature* **364**, 626 (1993).
25. J. P. Nowicki, D. Duvall, H. Poignet, B. Scatton, *Eur. J. Pharmacol.* **204**, 339 (1991).
26. J. S. Beckman, M. Carson, C. D. Smith, W. Koppenol, *Nature* **364**, 584 (1993).
27. A. Y. Chiu, E. W. Chen, S. Loera, *J. Comp. Neurol.* **328**, 351 (1993).
28. M. E. Gurney, H. Yamamoto, Y. Kwon, *J. Neurosci.* **12**, 3241 (1992).
29. We thank R. Huntress, S. Potter, and R. Hallewell for research materials and F. Cutting, B. Lom, and R. Mihalik for technical assistance. Supported by the National Institutes of Neurological Disorders and Stroke, the Muscular Dystrophy Association, the Les Turner ALS Foundation, the Searle Family Center for Neurological Disorders, the Vena E. Schaaf ALS Research Fund, and the Herbert and Florence C. Wenske Foundation.

24 March 1994; accepted 20 May 1994

Molecular Genetic Analyses of the Tyrolean Ice Man

Oliva Handt, Martin Richards, Marion Trommsdorff, Christian Kilger, Jaana Simanainen, Oleg Georgiev, Karin Bauer, Anne Stone, Robert Hedges, Walter Schaffner, Gerd Utermann, Bryan Sykes, Svante Pääbo*

An approximately 5000-year-old mummified human body was recently found in the Tyrolean Alps. The DNA from tissue samples of this Late Neolithic individual, the so-called "Ice Man," has been extracted and analyzed. The number of DNA molecules surviving in the tissue was on the order of 10 genome equivalents per gram of tissue, which meant that only multi-copy sequences could be analyzed. The degradation of the DNA made the enzymatic amplification of mitochondrial DNA fragments of more than 100 to 200 base pairs difficult. One DNA sequence of a hypervariable segment of the mitochondrial control region was determined independently in two different laboratories from internal samples of the body. This sequence showed that the mitochondrial type of the Ice Man fits into the genetic variation of contemporary Europeans and that it was most closely related to mitochondrial types determined from central and northern European populations.

In September 1991, the frozen mummy of a man was found in the Tyrolean Alps. Radiocarbon dates of skin and bone samples indicated an age between 5100 and 5300 years (*1*). Because no comparable archaeological discovery exists, this find has attracted considerable scientific and public interest. It has also been the subject of various rumors and even allegations of fraud (*2*). Molecular genetic investigations of the Ice Man could address some of the questions surrounding the find. Comparisons of DNA sequences from the body with contemporary populations may reveal aspects of his ethnic affiliation. Molecular studies of other organisms such as viruses or bacteria associated with the body may furthermore illuminate the evolution of these organisms. As a first step toward such investigations, we have analyzed the state of preservation of the DNA in the Ice Man and determined the sequence of a hypervariable segment of the mitochondrial control region from numerous samples removed from the body.

Ancient DNA has been retrieved from a variety of plant, animal, and human remains (*3*, *4*) that go back a few tens of thousands of years as well as from some fossils that are millions of years old (*5*–*7*), although the latter results are partially controversial (*8*). In most cases, work on archaeological DNA has been limited to mitochondrial DNA because its high copy number increases the chance of survival of a few molecules in the face of molecular damage that accumulates post mortem. Because the body of the Ice Man has been frozen with the exception of a short period after its discovery, its DNA may be preserved better than that of other finds. This unusual condition might allow nuclear markers such as microsatellites to be studied in addition to mitochondrial DNA and thus open several additional avenues of study.

A total of eight samples of muscle, connective tissue, and bone were removed under sterile conditions from the left hip region of the body, which had been damaged during salvage of the mummy. Additionally, parts of one sample that has been radiocarbon dated (*1*) were analyzed. Extracts of DNA were made from 10 to 200 mg of each sample by a silica-based method that is highly efficient in the retrieval of ancient DNA (*9*). Enzymatic amplifications from the mitochondrial control region were attempted. Because this region encodes no structural gene products and evolves faster than other parts of the mitochondrial genome, it is particularly suited for the reconstruction of the history of human popula-

O. Handt, C. Kilger, K. Bauer, A. Stone, S. Pääbo, Institute of Zoology, University of Munich, Postfach 202136, D-80021 München, Germany.
M. Richards and B. Sykes, Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DU, UK.
M. Trommsdorff and G. Utermann, Institute of Medical Biology and Genetics, University of Innsbruck, Schöpfstrasse 41, A-6020 Innsbruck, Austria.
J. Simanainen, O. Georgiev, W. Schaffner, Institut für Molekularbiologie II, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.
R. Hedges, Research Laboratory for Archaeology, University of Oxford, 6 Keble Road, Oxford OX1 3QJ, UK.
*To whom correspondence should be addressed.

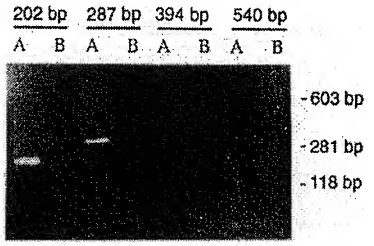


Fig. 1. Agarose gel electrophoresis of mitochondrial DNA amplification of different lengths from the Ice Man. For every primer pair, amplifications from (**A**) an extract of the Ice Man and (**B**) an extraction control are shown. The primer pairs used are as follows: L16055/H16218 (202 bp), L16055/H16303 (287 bp), L16055/H16410 (394 bp), and L15997/H16498 (540 bp), where L and H refer to the light and heavy strand, respectively, followed by the number to the nucleotide position (*14*) at the 3′ end of the primer. Migration positions of molecular size markers are given in numbers of base pairs.

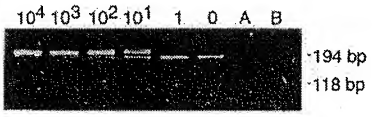


Fig. 2. Quantitation of mitochondrial DNA in an extract from the Ice Man. A dilution series of a competitor template, containing a 20-bp insertion in a mitochondrial fragment, was added to a constant amount of extract, and a PCR that used primers L16068/H16218 was performed as described in (*10*). The numbers above the lanes indicate the numbers of competition molecules added to the amplifications. (**A**) An extraction control and (**B**) a control where no template was added. Migration positions of molecular size markers are given in numbers of base pairs.

EXHIBIT 6

Focal loss of the glutamate transporter EAAT2 in a transgenic rat model of SOD1 mutant-mediated amyotrophic lateral sclerosis (ALS)

David S. Howland*†, Jian Liu‡, Yijin She*, Beth Goad*, Nicholas J. Maragakis§, Benjamin Kim§, Jamie Erickson*, John Kulik*, Lisa DeVito*, George Psaltis*, Louis J. DeGennaro*, Don W. Cleveland‡, and Jeffrey D. Rothstein§

*Department of Molecular Genetics, Wyeth Research, CN8000, Princeton, NJ 08543; ‡Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA 92093; and §Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD 21287

Edited by Thomas Maniatis, Harvard University, Cambridge, MA, and approved December 5, 2001 (received for review October 11, 2001)

Transgenic overexpression of Cu^{+2}/Zn^{+2} superoxide dismutase 1 (SOD1) harboring an amyotrophic lateral sclerosis (ALS)-linked familial genetic mutation ($SOD1^{G93A}$) in a Sprague–Dawley rat results in ALS-like motor neuron disease. Motor neuron disease in these rats depended on high levels of mutant SOD1 expression, increasing from 8-fold over endogenous SOD1 in the spinal cord of young presymptomatic rats to 16-fold in end-stage animals. Disease onset in these rats was early, ≈115 days, and disease progression was very rapid thereafter with affected rats reaching end stage on average within 11 days. Pathological abnormalities included vacuoles initially in the lumbar spinal cord and subsequently in more cervical areas, along with inclusion bodies that stained for SOD1, Hsp70, neurofilaments, and ubiquitin. Vacuolization and gliosis were evident before clinical onset of disease and before motor neuron death in the spinal cord and brainstem. Focal loss of the EAAT2 glutamate transporter in the ventral horn of the spinal cord coincided with gliosis, but appeared before motor neuron/axon degeneration. At end-stage disease, gliosis increased and EAAT2 loss in the ventral horn exceeded 90%, suggesting a role for this protein in the events leading to cell death in ALS. These transgenic rats provide a valuable resource to pursue experimentation and therapeutic development, currently difficult or impossible to perform with existing ALS transgenic mice.

Amyotrophic lateral sclerosis (ALS) is a late-onset neuromuscular disorder characterized by progressive motor dysfunction that leads to paralysis and eventually death. The pathology of the disease results from the death of large motor neurons in the spinal cord and brainstem (1, 2). ALS occurs in both sporadic and familial forms (3). Familial ALS accounts for ≈5–10% of all reported cases. Approximately 15–20% of familial ALS cases has been linked to inheritance in an autosomal dominant fashion of a mutant form of Cu^{+2}/Zn^{+2} superoxide dismutase 1 (SOD1) (4, 5). SOD1 normally functions in the regulation of oxidative stress by conversion of free radical superoxide anions to hydrogen peroxide and molecular oxygen. Over 90 distinct familial SOD1 mutations have been found to date. SOD1 mutations that have been tested in transgenic mice result in ALS-like motor neuron disease (6–8), but SOD1-null mice do not develop motor neuron disease (9). Furthermore, crossing SOD1-null mice with transgenic ALS mice does not alter disease onset or progression (10). Taken together, these results indicate that familial ALS does not result from loss of SOD1 function but rather an unidentified gain of function. There is no consensus as to the mechanism, and theories include alterations in SOD1 folding, oxidative stress from aberrant catalysis (11), or cytoplasmic aggregates (12). New studies also suggest that the disease is not cell autonomous—that nonneuronal cells are necessary for motor neuron degeneration (13, 14, ¶).

Transgenic mouse models expressing mutant forms of SOD1 (15–21) develop neuromuscular disease very similar to human ALS. Age of onset of disease varies as a function of both the type of mutant expressed in mouse and the relative expression levels attained. High expressing $SOD1^{G93A}$ (13-fold above endogenous SOD1) and G37R SOD1 (7–14-fold above endogenous SOD1) transgenic mice contain membrane-bound vacuoles in cell bodies (15, 22) and dendrites (15, 16, 22), which most likely result from degenerating mitochondria. Lower expressing $SOD1^{G93A}$ mice (7-fold above endogenous SOD1) also contain Lewy-body-like cytoplasmic inclusions in the cell bodies of motor neurons (21) containing SOD1, ubiquitin, and phosphorylated neurofilament (23). $SOD1^{G85R}$ transgenic mice expressing mutant SOD1 as little as 20% of endogenous levels also develop neuromuscular disease characterized by loss of large motor neurons in brainstem and in spinal cord (10, 17). No vacuolization has been reported in G85R mice or in similar mice expressing the murine counterpart mutation G86R (18, 24). However, these mice also develop cytoplasmic inclusions that appear in astrocytes and neurons before clinical signs of disease and dramatically increase in abundance with disease progression (10). $SOD1^{G85R}$ mice have also shown to be deficient in the spinal cord astroglial glutamate transporter EAAT2 (GLT-1), similar to observations in sporadic ALS (25), suggesting that astroglial dysfunction in ALS may contribute to motor neuron degeneration.

We sought to create a transgenic rat model for ALS by using mutant SOD1 to pursue experimental paradigms currently difficult or impossible to achieve in the smaller transgenic mouse models. Rats provide an advantage in pursuit of therapeutic strategies such as stem cell replacement and are the preferred laboratory animal species for pharmacological manipulations.

Materials and Methods

Generation and Characterization of Transgenic Rats Expressing Human $SOD1^{G93A}$. A 12-kb *Eco*RI/*Bam*HI restriction fragment of the human *SOD1* gene harboring the G93A mutation was microinjected into Sprague–Dawley rat embryos. Transgenic rats were produced as described (26). Embryos were allowed to develop to term and were analyzed for the presence of the transgene. Tail biopsies from 8-day-old rats were digested in proteinase K and then diluted 1:20 in dH_2O followed by heating at 95°C for 15 min. Two microliters were subjected to PCR by using primers SOD-i3f (5′-GTGGCATCAGCCCTAATCCA-3′) and SOD-E4r (5′-CACCAGTGTGCGGCCAATGA-3′) specific to human *SOD1* to determine the genotypes of founders and offspring.

Taqman quantitative DNA PCR was performed to determine DNA copy number of transgene loci segregating from the multiintegrant founders 26, 46, and 51 to their respective F1

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: SOD1, superoxide dismutase 1; ALS, amyotrophic lateral sclerosis; EMG, electromyography.

†To whom reprint requests should be addressed. E-mail: howland@war.wyeth.com.

¶Clement, A. M., Roberts, E. A., Goldstein, L. S. & Cleveland, D. W. (2001) *Soc. Neurosci. Abstr.* 27, no. 580.4.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.032539299

Table 1. Multiple transgene integrations in SOD1^{G93A} founders were resolved into individual lines

Line	Subline	dCt	Copy no.	Spinal cord hSOD1/rSOD1	Blood hSOD1/rSOD1	Pathology
26	26L	−3	8	nd	0.4	None
	26H	−6	64	8.6	0.8	~115 days
	26HL	−7	72	10.4	1.1	~102 days
46	46L	−0.5	1–2	0.4	<0.1	None
	46H	−2.5	5–6	2.6	0.2	None
51	51L	−2	4	2.2	0.5	None
	51H	−4	16	5.8	1.2	None
61	—	−2.3	4–5	2.4	0.1	None

ND, not determined; H, high copy; L, low copy; h, human; r, rat; dCt, delta cycle threshold.

generation progeny. Primer-probe sets specific for human *SOD1* and an internal normalizer gene, *Thy1.2*, were used in multiplex PCR on a Taqman 770 PCR thermocycler (PE Biosystems) following the manufacturer's recommended conditions. Data were represented as delta cycle threshold (dCt) and were converted to relative transgene DNA copy number by the equation $2^{(-dCt)}$ (Table 1).

Quantitation of SOD1 in Blood and SOD1 and EAAT2 in Spinal Cord. Blood samples from tail vein bleeds were solubilized in 10 vol of 50 mM Tris·HCl, pH 7.5/150 mM NaCl/5 mM EDTA/1% Nonidet P-40/1% SDS. For SOD1 detection, 25 μg was electrophoresed on 12% SDS/polyacrylamide gel and transferred to nitrocellulose. Cervical spinal cord was homogenized in 2 ml of 50 mM Tris·HCl, pH 7.5/150 mM NaCl/5 mM Na_2EDTA/1% Nonidet P-40/1% SDS, and 5 μg was electrophoresed as described above. For detection of EAAT2, ventral horn of cervical spinal cord was dissected by using 0.5-mm micropunches (Zivic–Miller) and homogenized as described above, and 25 μg of total protein was electrophoresed on 7.5% SDS/polyacrylamide gels. Western blots were probed with either anti-SOD1 (27) (1:5,000), anti-GLT-1 (EAAT2; 1:1,000; Chemicon), or anti-actin (C4; 1:10,000; Roche Molecular Biochemicals) Abs.

Immunohistochemical Analyses. Animals were killed by using approved animal welfare protocols and perfused by cardiac puncture with 4% paraformaldehyde/PBS. Muscle, brain, and spinal cord were removed followed by regional dissection of spinal cord and spinal nerve roots. Tissue blocks were embedded in paraffin or araldite for sectioning (7 and 1 μm, respectively). Immunostains and semithin plastic sections were processed as described (16, 17, 28). Hematoxylin and eosin stains of muscle and spinal cord were performed on paraffin sections, whereas semithin sections of spinal roots were stained with toluidine blue. Immunostaining was performed with Abs to neurofilament with SMI-32 (1:8,000; Sternberger–Mayer, Jarrettsville, MD), glutamate transporter GLT-1 (1:1,000), SOD1 (1:10,000), heat shock protein (HSP70; 1:100; StressGen Biotechnologies, Victoria, BC, Canada); ubiquitin (1:1,500; Dako), and glial fibrillary acidic protein (1:50; Dako).

Electrophysiological Recording. Electromyography (EMG) and nerve conductions were performed by using an ADI (Greenwich, CT) Powerlab/8SP stimulator and BioAMP amplifier followed by computer assisted data analysis (CHART 4.0 and SCOPE 3.5.6; ADI). Compound muscle action potentials were recorded by stimulating the sciatic nerve at the sciatic notch and recording from the foot. EMG was performed by using a bipolar needle and sampling at 200 Hz.

Results

Multiple Transgenic Rat Lines Express Mutant SOD1^{G93A}. We identified three SOD1^{G93A} founders that expressed mutant human SOD1 in blood (Fig. 1*A*). A fourth (founder 46) showed no detectable SOD1^{G93A} expression; however, subsequent immunoblots of whole blood from the F1 animals of this line did indeed show low-level SOD1^{G93A} (not shown). These four founders were bred to the F1 generation to establish transgenic lines.

Transgene transmission frequency to the F1 generation was greater than the expected 50% in lines 26, 46, and 51 and was determined to be the result of multiple transgene integration sites in each of these founders. Distinct transgene integrations can be resolved by using quantitative Taqman PCR if the number of transgene copies differs at each chromosomal site. This was indeed the case for lines 26, 46, and 51. Taqman PCR data were used to track inheritance of distinct low- or high-copy transgene loci by F1 generation animals thereby allowing us to establish separate sublincages for each of these lines (Table 1).

Development of Motor Neuron Disease in SOD1^{G93A} Transgenic Rats. SOD1^{G93A} founder 26 developed motor neuron disease at 93 days of age, whereas all other founders did not develop disease. Because of the multiple integration of the transgene, the F1 generation animals derived from founder 26 inherited either the high- or low-copy transgene locus or both (see Table 1). F1 animals containing only the low-copy locus (L26L) did not develop motor neuron disease. F1 animals that inherited both loci from founder 26 (L26HL) developed motor neuron disease by 93 days of age, the same age as disease onset in the founder. F1 animals that inherited only the high-copy locus (L26H) developed motor neuron disease between 104–121 days of age. The apparent earlier onset in L26HL vs. L26H animals most likely was the result of slightly higher mutant SOD1 expression (Table 1). Because the single high-copy locus (L26H) in rats was sufficient to elicit motor neuron disease, we chose to breed this subline to the F2 and subsequent generations for further analysis.

Mutant SOD1 Expression in SOD1^{G93A} Transgenic Rats. SOD1^{G93A} expression in the spinal cord of L26H transgenic rats was determined to be ≈8-fold above endogenous SOD1 as assessed by immunoblot analysis of young presymptomatic animals (Fig. 1*C*; Table 1). As expected, these levels exceeded other transgenic rat lines that did not go on to develop motor neuron disease (Table 1). SOD1^{G93A} expression in L26H rats was also evident across many brain regions as well as peripheral tissues (Fig. 1*B*), similar to that seen in described SOD1 transgenic mice (16). By end stage, mutant SOD1 levels accumulated ≈16-fold over endogenous, representing a further 2-fold increase in SOD1^{G93A} compared with levels in young presymptomatic rats (6 weeks old) (Fig. 1*C*). Spinal cord SOD1^{G93A} levels were directly compared

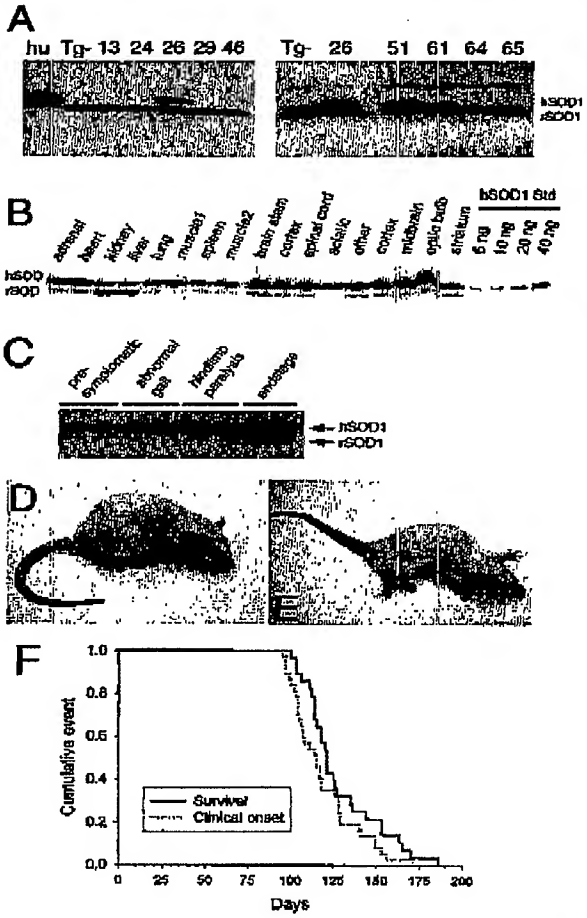


Fig. 1. Mutant SOD1 expression and disease in SOD1^{G93A} transgenic rats. SOD1 expression in blood from transgenic founders (*A*) is highest in founder number 26. L26H F1 generation rats exhibit SOD1^{G93A} expression throughout the nervous system and peripheral tissues (*B*). SOD1^{G93A} expressed at ≈8-fold over endogenous in young (6 weeks) presymptomatic transgenic rat spinal cord increases to ≈16-fold by end-stage disease (16 weeks) (*C*). Normal age-matched littermate control animal (*D*) at ≈120 days compared with an end-stage transgenic rat showing signs of muscle wasting, paralysis of both hindlimbs and one forelimb (*E*). Kaplan–Meyer survival curve (n = 25) generated from F2 generation L26H transgenic rats depicting disease onset and survival.

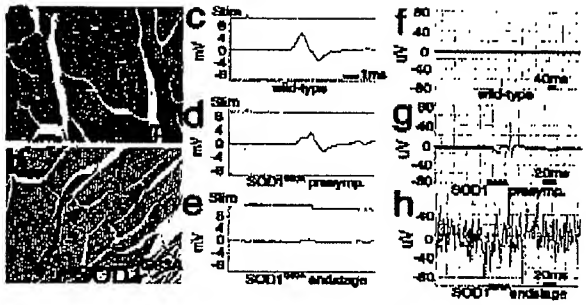


Fig. 2. Muscle atrophy and denervation in SOD1^{G93A} rats. Leg muscle myofibers from end-stage (age >120 days) SOD1^{G93A} rats were often seen as groups of atrophic angular fibers (*b*, arrows), compared with aged-matched control rats (*a*). Compound muscle action potential in nontransgenic control foot muscle (*c*; 5.48 mV) was reduced in the foot (*d*; 4.3 mV) in presymptomatic rats and was almost unobtainable in end-stage foot (*e*; 0.71 mV) after supramaximal stimulation (1 ms per division). Needle EMG of presymptomatic SOD1^{G93A} rat (*g*) demonstrates a rare fibrillation potential recorded in the lumbosacral paraspinous muscles compared with age-matched wild-type control rat (*f*). EMG of end-stage (>125 days age) SOD1^{G93A} rat (*h*) revealed continuous fibrillation potentials and positive sharp waves (20 ms per division).

in end-stage L26H transgenic rats and the previously described G1H and G1L transgenic mice, which also express SOD1^{G93A}. We found that SOD1^{G93A} levels in end-stage G1H and G1L (15, 21, 22) transgenic mice were 3- and 1.5-fold, respectively, higher than levels attained in end-stage L26H transgenic rats (data not shown).

Characterization of Motor Neuron Disease in SOD1^{G93A} L26H Transgenic Rats. A subset (n = 25) of F2 generation animals for L26H were observed closely for onset of disease symptoms, as well as progression to death. Onset of motor neuron disease was scored as the first observation of an abnormal gait or evidence of hindlimb weakness. Affected animals were tested daily for the ability to right themselves after being turned on either side for a maximum of 30 sec; failure at this task was seen in end-stage animals and scored as "death" (see Fig. 1*E*). All end-stage animals were killed. Righting reflex failure was coincident with complete paralysis of both hindlimbs and at least 1 forelimb. F2 L26H transgenic rats had an average age of onset of motor neuron disease of 115 days. Onset typically appeared as hindlimb abnormal gait and progressed very quickly (1–2 days) to overt hindlimb paralysis, typically affecting one limb first. Within 1–2 days, the second hindlimb was involved, although animals could still ambulate through the use of forepaws. Affected rats showed signs of weight loss, poor grooming, and porphyrin staining around the eyes. L26H transgenic rats typically reached end-stage disease very quickly, an average of 11 days after onset of symptoms. All F2 generation L26H transgenic rats monitored for this study reached end-stage disease within 173 days after birth.

Muscle Pathology and Impaired Function in SOD1^{G93A} L26H Rats. Leg muscles (distal and proximal) from end-stage rats revealed obvious and frequent angular atrophic myofibers, most often in discrete clumps typical of neurogenic atrophy (Fig. 2*b*), whereas muscles from wild-type littermate controls were normal (Fig. 2*a*). In parallel, electrophysiologic recordings from end-stage SOD1^{G93A} rats (n = 5) exhibiting obvious hindlimb paralysis or paresis revealed markedly reduced amplitude of compound motor action potentials (CMAP) in the intrinsic foot muscles (Fig. 2*e*), indicating motor neuron loss (compare 5.48 mV in wild-type animals to 0.71 mV in end-stage SOD1^{G93A} rats). CMAP in presymptomatic animals was diminished only partially (Fig. 2*d*; n = 5) compared with littermate controls (Fig. 2*c*). Needle EMG of age-matched wild-type rats demonstrated absence of any spontaneous activity, compared with rare fibrillation potentials in paraspinal muscles from presymptomatic rats (Fig. 2 *f* and *g*). Continuous fibrillation potentials and positive sharp waves were evident in leg muscles from end-stage L26H rats (Fig. 2*h*).

Immunohistochemical Characterization of SOD1^{G93A} L26H Transgenic Rats. Analysis of hematoxylin/eosin-stained sections of lumbar spinal cord from end-stage SOD1^{G93A} rats (n = 10) revealed a

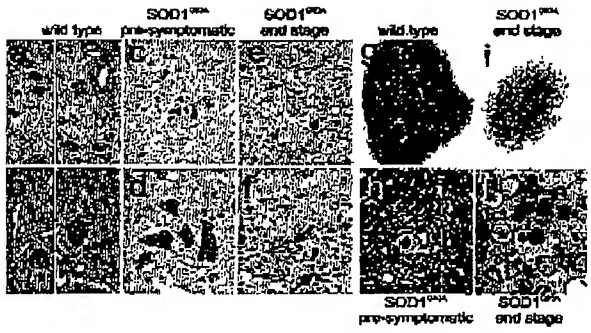


Fig. 3. Motor neuron and axon loss in SOD1^{G93A} rats. Ventral spinal cord gray matter reveals vacuolar degeneration in the neuropil of presymptomatic SOD1^{G93A} rats (*c* and *d*) and astrogliosis and loss of motor neurons in end-stage rats (*e* and *f*) compared with age-matched wild-type rats (*a* and *b*). Glial nodules, around remnants of degenerating motor neurons, were evident throughout the ventral gray matter (*f*, arrows). Ventral motor roots from an end-stage rat (*i*) were atrophic compared with aged-matched control roots (*g*). Closer inspection revealed active ongoing degeneration in end-stage SOD1^{G93A} ventral roots (*j*), whereas roots from presymptomatic rats showed little degeneration (*h*). Magnification: ×4, *g* and *i*; ×10, *a*, *c*, and *e*; ×40, *b*, *d*, and *f*; ×100, *h* and *j*.

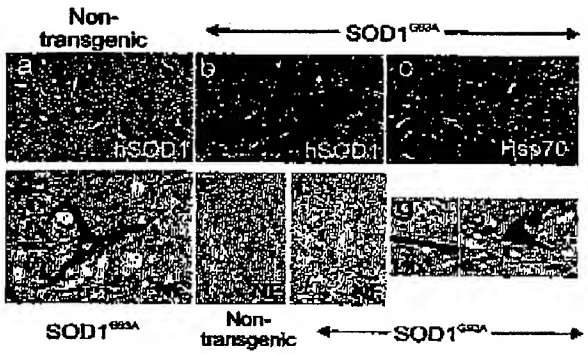


Fig. 4. Aberrant accumulations of proteins in SOD1^{G93A} rats. Accumulation in the neuropil of SOD1^{G93A} rats (*b*) compared with age-matched wild-type rats (*a*). Hsc70 (*c*) and ubiquitin (Ub) (*g*) were abnormally accumulated in the neuropil and cytoplasm of ventral gray neurons. Similarly, neurofilament (NF) aggregates were found in the soma of large motor neurons (*d*) and their axons, often in spheroid structures in the neuropil and especially in the ventral root zone white matter (*f*) compared with dorsal white tracts (*e*).

dense gliosis with a complete loss of ventral large motor neurons (alpha-motor neurons) as shown in Fig. 3 *e* and *f* compared with similarly aged wild-type rats (Fig. 3 *a* and *b*). Closer inspection demonstrated frequent "glial nodules" (Fig. 3*f*, arrows), representing active degeneration and engulfment of neurons. Inspection of ventral horn gray matter from lumbar spinal cord from presymptomatic rats (≈90–100 days of age) revealed a normal population of motor neurons but a profound vacuolar degeneration of the neuropil (Fig. 3 *c* and *d*), similar to that seen in end-stage SOD1^{G93A} mice (15, 22). However, in the rat these vacuoles were transient, appearing at the time of active motor neuron loss but were nearly absent in the lumbar cord by end-stage disease (Fig. 3 *e* and *f*). Brainstem and cervical spinal cord of end-stage rats also revealed vacuolar and glial nodule changes in motor neurons (not shown), albeit these appeared later in these regions, again consistent with vacuolar presence preceding neuronal loss. In concert with the changes in gray matter, ventral roots from end-stage SOD1^{G93A} rats were atrophic (Fig. 3*i*). On closer inspection (Fig. 3*j*) active degeneration of most axons was observed with macrophage infiltration and myelin ovoids. In contrast, analysis from presymptomatic SOD1^{G93A} rat ventral roots ($n = 5$) showed almost normal-appearing axons (Fig. 3*h*), compared with age-matched controls (Fig. 3*g*), with rare (1–2 axons per root) undergoing degeneration (Fig. 3*h* arrow).

As was reported in earlier examples of SOD1 mutant-mediated disease in mice (10), onset of clinical disease was accompanied by aggregates of SOD1 throughout the rat ventral horn (Fig. 4*b*) and brain (not shown) especially in prominent focal deposits in which mutant SOD1 immunoreactivity was frequently most robust at the perimeter. Similar pathology was not found in nontransgenic controls (Fig. 4*a*). These aggregates could be found in a few surviving motor neuron perikarya, axons, and surrounding glia. Aggregates were intensely stained with Abs to ubiquitin (Fig. 4 *g*), consistent with disruption in protein clearance by the proteosome. Aggregates also contained endogenous Hsc70 especially within surviving motor neuron cell bodies (Fig. 4*c*), suggesting mutant-dependent depletion of the intracellular protein folding chaperone pool. Aberrant accumulations of neurofilaments, reported in SOD1^{G93A}-expressing transgenic mice (29), were prominent abnormalities after disease onset, both in perikarya (Fig. 4*d*) and in distended axonal swellings [compare the neurofilament staining in transgenic axons (Fig. 4*f*) with that of the wild-type littermate controls (Fig. 4*e*)]. These accumulations were selective for axons within the ventral root exit zone and were not found in the dorsal ascending columns (not shown).

EAAT2 Deficits in the Ventral Horn Spinal Cord of SOD1^{G93A} L26H Rats. EAAT2 is the predominant glutamate transporter in the central nervous system, normally expressed widely throughout the spinal gray matter (Fig. 5*a*) in astrocytes but not in motor neurons (arrows, Fig. 5*b*). Previous studies have documented a profound loss of the protein in sporadic and familial ALS (25, 28, 30). In presymptomatic SOD1^{G93A} rats, just before disease onset, motor neurons are still present (Fig. 5*c*, arrows) and ventral horns have not started to degenerate (Fig. 3*h*). At this time point, there is an obvious patchy loss of EAAT2 immunoreactivity in the ventral horn (Fig. 5*c*). By end stage, there is a profound focal loss of EAAT2 immunoreactivity despite a striking increase in the number of astrocytes (Fig. 5 *d* and *e*). These changes were mirrored by a quantitative loss of EAAT2 immunoreactivity measured from immunoblots of extracts from spinal cord, especially in the ventral gray regions (Fig. 5*f*). Assays of glutamate transport also confirmed a nearly 50% loss of functional transport (data not shown). Astroglial reactivity, as revealed by glial fibrillary acidic protein immunostaining, also showed activation before motor neuron degeneration, in presymptomatic spinal cord ventral gray (Fig. 5*h*) compared with nontransgenic controls (Fig. 5*g*), followed by a more dramatic activation (Fig. 5 *i* and *j*) in end-stage tissue.

Discussion

We have generated transgenic Sprague–Dawley rats that express human mutant SOD1^{G93A} at levels ≈8-fold over endogenous SOD1 in the spinal cord of young presymptomatic rats. This level of expression was sufficient to cause an ALS-like motor neuron disease in rats by 3–4 months of age. Additional transgenic lines expressing mutant SOD1 between 0.1- to ≈6-fold over endogenous levels of SOD1 have not developed any signs of motor neuron disease by 1 year of age. Recapitulation of an ALS-like motor neuron disease in the transgenic rat using the G93A

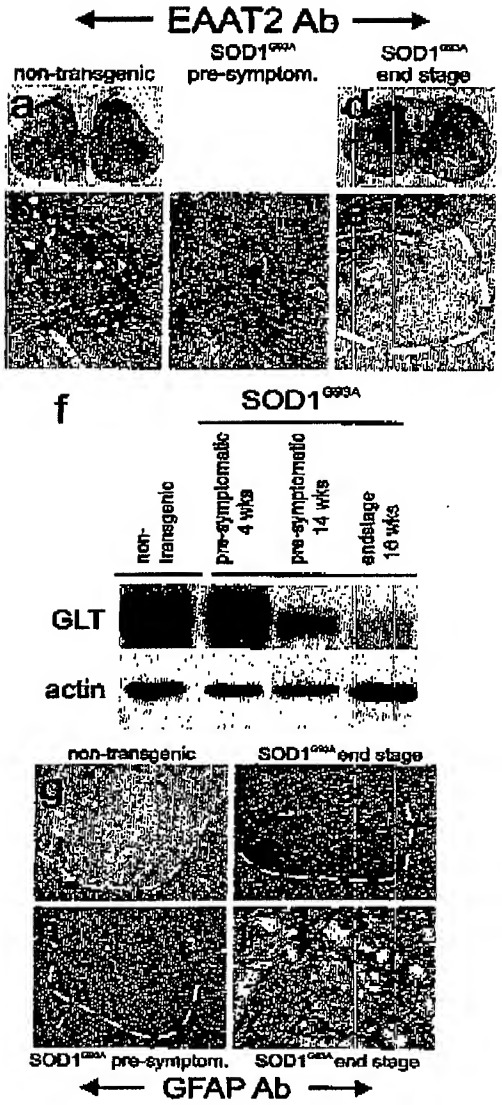


Fig. 5. Astroglial alterations in SOD1^{G93A} rats. The usual ubiquitous astroglial expression of the glutamate transporter EAAT2 (*a* and *b*), surrounding motor neurons (arrows), was markedly altered in SOD1^{G93A} rats with a patchy loss in the ventral horn in presymptomatic rats (*c*) and almost a complete loss of protein in end-stage ventral gray from SOD1^{G93A} rats (*d* and *e*). This loss of EAAT2 (GLT) was paralleled in immunoblots from ventral gray of presymptomatic and end-stage rats (*f*). In parallel, astroglial expression of glial fibrillary acidic protein (GFAP) increased somewhat in presymptomatic ventral gray (*h*), compared with age-matched wild-type control (*g*), and was markedly increased in end-stage rats (*i*), especially around rare motor neuron profiles (*j*).

mutant SOD1 clearly depended on the ability to obtain high-level transgene expression in the spinal cord as reported for the SOD1^{G93A} (15) and SOD1^{G37R} (16) transgenic mice.

No overt motor neuron loss was evident in presymptomatic SOD1^{G93A} transgenic rats between 3–4 months of age as determined by both histological and electrophysiological observations. However, we noted the appearance of vacuoles in motor neurons as well as gliosis preceding both motor neuron loss and clinical signs of disease in rats. The presence of vacuoles was transient, correlating with the time of active motor neuron loss. In the most affected regions vacuoles were nearly absent by end-stage disease. Progression to end-stage paralysis was rapid, with an average of 11 days after first observation of symptoms. This finding is in contrast to the slower progression of disease observed in SOD1^{G93A} transgenic mice (G1H and G1L) where disease duration approached 60–70 days (15, 21, 22) but instead was more similar to that reported for SOD1^{G85R} mice (17) whose disease duration was only 7–14 days. Mutant SOD1 levels in end-stage G1L and G1H transgenic mouse spinal cord (15, 21, 22) were higher than in SOD1^{G93A} L26H transgenic rats, and therefore the rapid progression of disease in the SOD1^{G93A} transgenic rats seems not to be a function of expression levels but rather may be characteristic of a species difference in the presentation of clinical phenotype. The rapid decline of the SOD1^{G93A} rats coincided with substantial loss of spinal cord motor neurons as well as marked increases in gliosis and degeneration of muscle integrity and function.

The astroglial glutamate transporter EAAT2 is the primary means of maintaining low extracellular glutamate levels. Loss of this protein induced by either pharmacological or molecular methods *in vitro* and *in vivo* results in increased extracellular glutamate, as measured by microdialysis and excitotoxic neuronal degeneration, including degeneration of motor neurons. Elevations of extracellular glutamate and loss of EAAT2 are characteristic of at least 40% of sporadic patients with ALS, and similar changes have been observed in the mutant mouse models of the disease (17, 25, 31–33). Interestingly, a recent study of a similar transgenic rat model, however, did not observe changes in cerebrospinal fluid (CSF) glutamate (34). The reason for the difference between that rat model, the work in the current study, and previous human observations is not clear. However, a focal loss of EAAT2 would be expected to increase glutamate only locally and therefore might not be detectable in the CSF. In addition, CSF glutamate measurements are fraught with technical problems.

The cause of EAAT2 loss is not known, but multiple studies demonstrate that astroglial changes can occur early, before actual motor neuron degeneration (13, 17). However, loss of neurons can lead to glial responses that include transient down-regulation of EAAT2 expression (35, 36). Yet, there is no loss of EAAT2 in another motor neuron disease, spinal muscular atrophy (33, 37). Previous reports have documented a loss of EAAT2 to ≈50% its normal level in SOD1^{G85R} transgenic mice (17) by using whole spinal cord at end-stage disease. The current study provides a thorough evaluation of EAAT2 at a time point when motor neurons are intact histologically and physiologically, as revealed by EMG/nerve conduction studies. At these "early" time points, there was a patchy loss of EAAT2 expression around motor neurons in the ventral gray areas of the spinal cord, suggesting that the loss of EAAT2 may contribute to motor neuron degeneration. Concomitant with decreased EAAT2 expression was a marked increase in gliosis, and by end stage, where motor neuron loss is severe, EAAT2 was present at only 5–10% of normal levels in the ventral horn. Importantly, the contribution of altered EAAT2 expression to neuronal death/injury was demonstrated by a recent study where EAAT2 overexpression offered protection in SOD1^{G93A} mice.[‖]

[‖]Sutherland, M. L., Martinowich, K. & Rothstein, J. D. (2001) *Soc. Neurosci. Abstr.* 27, no. 807.6.

We describe a transgenic rat model for ALS based on the *SOD1*^{G93A} mutation. The clinical and pathological changes displayed resemble the "high expressing" SOD1^{G93A} mice first described by Gurney *et al.* (15) including a characteristic vacuolar degeneration of the neuropil, which seems to occur just before motor neuron degeneration and aggregates staining with SOD1 and neurofilament. Proteins, Hsc70 and ubiquitin, involved in protein folding as well as degradation are also present in these aggregates in these transgenic rats. Notable differences between the rat and mouse models, however, include a more rapid progression of disease and the transient appearance of vacuoles in the transgenic rat. The rapid decline of the SOD1^{G93A} rats to end stage could account for the disappearance of vacuoles in sections of the spinal cord that display severe motor neuron loss.

We have also demonstrated, using the SOD1^{G93A} rats, that EAAT2 levels decrease in the spinal cord before clinical onset of symptoms and that decrease becomes more severe by end-stage sickness, suggesting a role for glutamate toxicity and astroglial dysfunction in disease pathogenesis.

D.S.H. thanks Drs. Lucie Bruijn, John Moyer, Seung Kwak, and Erika Holzbaur for critical comments and advice. D.W.C. and J.D.R. gratefully acknowledge support from National Institutes of Health Grants NS 27036, NS33958, and AG 12992, and the Center for ALS Research. J.L. is the recipient of a fellowship from the Spinal Cord Disease Foundation. D.W.C. receives salary support from the Ludwig Institute for Cancer Research. This work was initiated by the ALS Association as part of its Lou Gehrig Challenge Initiative.

1. Dellalle, M. B. & Carpenter, S. (1984) *J. Neurol. Sci.* **63**, 241–250.
2. Bunker, B. Q. (1985) In *Myology*, eds. Engel, A. G. & Banker, B. Q. (McGraw-Hill, New York), pp. 2031–2066.
3. Horton, W. A., Eldredge, R. & Brody, J. A. (1976) *Neurology* **26**, 460–465.
4. Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J. P., Deng, H. X. *et al.* (1993) *Nature (London)* **362**, 59–62.
5. Deng, H. X., Hentati, A., Tainer, J. A., Iqbal, Z., Cayabyab, A., Hung, W. Y., Getzoff, E. D., Hu, P., Herzfeldt, B., Roos, R. P., *et al.* (1993) *Science* **261**, 1047–1051.
6. Morrison, B. M., Morrison, J. H. & Gordon, J. W. (1998) *J. Exp. Zool.* **282**, 32–47.
7. Wong, P. C., Rothstein, J. D. & Price, D. L. (1998) *Curr. Opin. Neurobiol.* **8**, 791–799.
8. Shibata, N. (2001) *Neuropathology* **21**, 82–92.
9. Reaume, A. G., Elliot, J. L., Hoffman, E. K., Kowall, N. W., Ferrante, R. J., Siwek, D. F., Wilcox, H. M., Flood, D. G., Beal, M. F., Brown, R. H., *et al.* (1996) *Nat. Genet.* **13**, 43–47.
10. Bruijn, L. I., Houseweart, M. K., Kato, S., Anderson, K. L., Anderson, S. D., Ohama, E., Reaume, A. G., Scott, R. W. & Cleveland, D. W. (1998) *Science* **281**, 1851–1854.
11. Beckman, J. S., Carson, M., Smith, C. D. & Koppenol, W. H. (1993) *Nature (London)* **364**, 584.
12. Johnston, J. A., Dalton, M. J., Gurney, M. E. & Kopito, R. R. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 12571–12576.
13. Gong, Y. H., Parsadanian, A. S., Andreeva, A., Snider, W. D. & Elliott, J. L. (2000) *J. Neurosci.* **20**, 660–665.
14. Pramatarova, A., Laganiere, J., Roussel, J., Brisebois, K. & Rouleau, G. A. (2001) *J. Neurosci.* **21**, 3369–3374.
15. Gurney, M. E., Pu, H., Chiu, A. Y., Dal Canto, M. C., Polchow, C. Y., Alexander, D. D., Caliendo, J., Hentati, A., Kwon, Y. W. & Deng, H. X. (1994) *Science* **264**, 1772–1775.
16. Wong, P. C., Pardo, C. A., Borchelt, D. R., Lee, M. K., Copeland, N. G., Jenkins, N. A., Sisodia, S. S., Cleveland, D. W. & Price, D. L. (1995) *Neuron* **14**, 1105–1116.
17. Bruijn, L. I., Becher, M. W., Lee, M. K., Anderson, K. L., Jenkins, N. A., Copeland, N. G., Sisodia, S. S., Rothstein, J. D., Borchelt, D. R., Price, D. L. & Cleveland, D. W. (1997) *Neuron* **18**, 327–338.
18. Ripps, M. E., Huntley, G. W., Hof, P. R., Morrison, J. H. & Gordon, J. W. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 689–693.
19. Brannstrom, T., Ernhill, K., Jonsson, A., Nilsson, A. & Marklund, S. (2000) *Brain Pathol.* **10**, 775.
20. Friedlander, R., Brown, R., Gagliardini, V., Wang, J. & Yuan, J. (1997) *Nature (London)* **388**, 31.
21. Dal Canto, M. & Gurney, M. (1997) *Acta. Neuropathol.* **93**, 537–550.
22. Dal Canto, M. & Gurney, M. E. (1995) *Brain Res.* **676**, 25–40.
23. Shibata, N., Hirano, A., Kobayashi, M., Dal Canto, M. C., Gurney, M. E., Komori, T., Umahara, T. & Asayama, K. (1998) *Acta Neuropathol.* **95**, 136–142.
24. Morrison, B. M., Janssen, W. G., Gordon, J. W. & Morrison, J. H. (1996) *J. Comp. Neurol.* **373**, 619–631.
25. Rothstein, J. D., Van Kammen, M., Levey, A. I., Martin, L. J. & Kuncl, R. W. (1995) *Ann. Neurol.* **38**, 73–84.
26. Hogan, B. (1983) *Nature (London)* **306**, 313–314.
27. Pardo, C. A., Xu, Z., Borchelt, D. R., Price, D. L., Sisodia, S. S. & Cleveland, D. W. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 954–958.
28. Rothstein, J. D., Martin, L., Levey, A. I., Dykes-Hoberg, M., Jin, L., Wu, D., Nash, N. & Kuncl, R. W. (1994) *Neuron* **13**, 713–725.
29. Tu, P. H., Raju, P., Robinson, K. A., Gurney, M. E., Trojanowski, J. Q. & Lee, V. M. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3155–3160.
30. Rothstein, J. D., Martin, L. J. & Kuncl, R. W. (1992) *N. Engl. J. Med.* **326**, 1464–1468.
31. Guo, Z., Kindy, M. S., Kruman, I. & Mattson, M. P. (2000) *J. Cereb. Blood Flow Metab.* **20**, 463–468.
32. Pedersen, W. A., Fu, W., Keller, J. N., Markesbery, W. R., Appel, S., Smith, R. G., Kazarskis, E. & Mattson, M. P. (1998) *Ann. Neurol.* **44**, 819–824.
33. Shaw, P. J., Chinnery, R. M. & Ince, P. G. (1994) *Brain Res.* **655**, 195–201.
34. Nagai, M., Aoki, M., Miyoshi, I., Kato, M., Pasinelli, P., Kasai, N., Brown, R. H. & Itoyama, Y. (2001) *J. Neurosci.* **21**, 9246–9254.
35. Ginsberg, S. D., Martin, L. J. & Rothstein, J. D. (1995) *J. Neurochem.* **65**, 2800–2803.
36. Ginsberg, S. D., Rothstein, J. D., Price, D. L. & Martin, L. J. (1996) *J. Neurochem.* **67**, 1208–1216.
37. Lin, C. L., Bristol, L. A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L. & Rothstein, J. D. (1998) *Neuron* **20**, 589–602.

EXHIBIT 7

Human Molecular Genetics, 2003, Vol. 12, No. 6 617–624
DOI: 10.1093/hmg/ddg075

Transgenic rat model of Huntington's disease

Stephan von Hörsten[1,*], Ina Schmitt[2,†], Huu Phuc Nguyen[1], Carsten Holzmann[3], Thorsten Schmidt[4], Thomas Walther[5], Michael Bader[6], Reinhard Pabst[1], Philipp Kobbe[1], Jana Krotova[1], Detlef Stiller[7], Ants Kask[8], Annika Vaarmann[8], Silvia Rathke-Hartlieb[9], Jörg B. Schulz[9], Ute Grasshoff[4], Ingrid Bauer[3], Ana Maria Menezes Vieira-Saecker[2,†], Martin Paul[10], Lesley Jones[11], Katrin S. Lindenberg[12], Bernhard Landwehrmeyer[12], Andreas Bauer[13], Xiao-Jiang Li[14] and Olaf Riess[4]

[1]Department of Functional and Applied Anatomy, Hannover Medical School, Hannover, Germany, [2]Department of Molecular Human Genetics, University of Bochum, Bochum, Germany, [3]Department of Medical Genetics, Children's Hospital, University of Rostock, Rostock, Germany, [4]Department of Medical Genetics, University of Tübingen, Tübingen, Germany, [5]Department of Cardiology, University Hospital Benjamin Franklin, Free University of Berlin, Berlin, Germany, [6]Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany, [7]Department of Non-invasive Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany, [8]Department of Pharmacology, University of Tartu, Tartu, Estonia, [9]Department of Neurology, University of Tübingen, Tübingen, Germany, [10]Institute of Pharmacology and Toxicology, University Hospital Benjamin Franklin, Free University of Berlin, Berlin, Germany, [11]Institute of Medical Genetics, University of Wales, College of Medicine, Cardiff, UK, [12]Department of Neurology, University of Ulm, Ulm, Germany, [13]Institute of Medicine, Research Center, Jülich, Germany and [14]Department of Human Genetics, Emory University School of Medicine, Atlanta, CA, USA

Received October 24, 2002; Revised and Accepted January 4, 2003

Huntington's disease (HD) is a late manifesting neurodegenerative disorder in humans caused by an expansion of a CAG trinucleotide repeat of more than 39 units in a gene of unknown function. Several mouse models have been reported which show rapid progression of a phenotype leading to death within 3–5 months (transgenic models) resembling the rare juvenile course of HD (Westphal variant) or which do not present with any symptoms (knock-in mice). Owing to the small size of the brain, mice are not suitable for repetitive *in vivo* imaging studies. Also, rapid progression of the disease in the transgenic models limits their usefulness for neurotransplantation. We therefore generated a rat model transgenic of HD, which carries a truncated huntingtin cDNA fragment with 51 CAG repeats under control of the native rat huntingtin promoter. This is the first transgenic rat model of a neurodegenerative disorder of the brain. These rats exhibit adult-onset neurological phenotypes with reduced anxiety, cognitive impairments, and slowly progressive motor dysfunction as well as typical histopathological alterations in the form of neuronal nuclear inclusions in the brain. As in HD patients, *in vivo* imaging demonstrates striatal shrinkage in magnetic resonance images and a reduced brain glucose metabolism in high-resolution fluor-deoxy-glucose positron emission tomography studies. This model allows longitudinal *in vivo* imaging studies and is therefore ideally suited for the evaluation of novel therapeutic approaches such as neurotransplantation.

INTRODUCTION

Huntington's disease (HD) is an autosomal dominant disorder caused by an expanded and unstable CAG trinucleotide repeat within the coding region of the HD gene (IT15) (1). The mutation leads to a progressive degeneration of neurons primarily in striatum and cerebral cortex. Clinically, HD is characterized by movement abnormalities, cognitive impairments, and emotional

*To whom correspondence should be addressed at: Department of Functional and Applied Anatomy, OE 4120, Hannover Medical School, Carl Neuberg Str. 1, 30625 Hannover, Germany. Tel: +49 5115322868, Fax: +49 5115328868; Email: hoersten.stephan.von@mh-hannover.de
†Present address: Department of Neurology, University of Bonn, Bonn, Germany.

disturbances (2). In general, movement disturbances begin with chorea. Depressed mood and more subtle deficits apparent in neuropsychological tests may precede motor symptoms by years. The disease progresses relentlessly until death within 15–20 years. No effective treatment to influence the onset or the progression is presently available.

Many attempts have been made to generate animal models of HD. Excitotoxin models replicate many of the histological and neurochemical features as well as some of the motor and cognitive signs of HD (3–5), but neurodegeneration is not truly progressive. Therefore, their usefulness for the evaluation of treatment effects is limited.

Transgenic animal models of HD (6–11) provide new ways of studying the neuropathological mechanisms underlying HD. In particular the R6/2 transgenic mouse line, which expresses the first exon of the human HD gene carrying 141–157 CAG repeat expansions (6), develops a number of key features of HD, including progressive motor deterioration (12,13), appearance of neuronal intranuclear inclusions (14), discriminative learning impairments (15), and altered emotionality (16). However, R6/2 mice express very large numbers of CAG repeats that are only found in the juvenile type of HD. A rapid disease progression associated with diabetes in R6/2 mice (13) is not typical for the adult-type HD and may complicate the assessment of potential therapeutic approaches. Although HD transgenic mice provide important insights into the molecular basis of HD, there is still a need for animal models which resemble the common adult type of disease and which are more suitable for repetitive *in vivo* imaging. These rapidly emerging techniques offer the opportunity to compare directly the pathological alterations of the human condition with the corresponding animal model in longitudinal studies (17).

In this report, we describe the first transgenic rat model bearing a human HD mutation with a high-end adult onset allele of 51 CAG repeats that exhibits progressive neurological, neuropathological and neurochemical phenotypes closely resembling the common late manifesting and slowly progressing type of disease. We demonstrate that HD transgenic rats are well suited for complex behavioral studies and the evaluation of *in vivo* progression markers using high-resolution PET and MRI.

RESULTS

Generation of the HD transgenic rat model

A 1962 bp rat HD cDNA fragment (18) carrying expansions of 51 CAG repeats under the control of 885 bp of the endogenous rat HD promotor (19) was used for microinjection (Fig. 1A). Two founders were obtained and transgenic lines established. Of these, we followed up line 2762 for more than 2 years and found the CAG repeat length remaining stable in more than 147 meioses (data not shown). The mutant amino terminal portion of huntingtin is expressed in the brain as shown by western blot analysis (Fig. 1B), in particular in the frontal and temporal cortex, the hippocampus, the basal ganglia, and the mesencephalon, but at a much lower level in the cerebellum or the spinal cord (Fig. 1C).

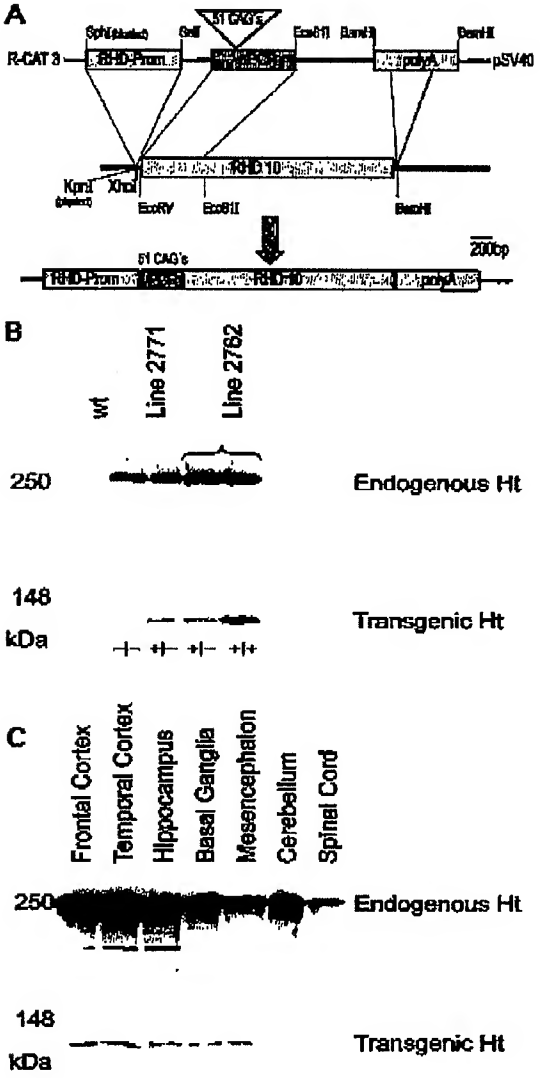


Figure 1. Transgene construct and huntingtin expression in transgenic rats. (A) The first 154 bp of a partial huntingtin cDNA spanning 1962 bp of the N-terminal rat sequence (RHD10) (18) were replaced by a PCR product from the affected allele of a HD patient. The cDNA is driven by a 885 bp fragment of the rat HD promotor (position −900 to −15 bp) (19). A 200 bp fragment containing the SV40 polyadenylation signal was finally added downstream of the cDNA resulting in RHD/Prom51A. (B) Western blot analysis of brain tissue of transgenic rat line 2771 and 2762 using polyclonal anti-huntingtin antibody 675 demonstrates a 75 kDa product representing the expression of the transgene although at a lower level than the endogenous protein. Homozygotic rats (+/+) express about double the amount of the transgene protein as hemizygotic lines (+/−). (C) Western blot analysis of tissue from different brain areas of transgenic rat line 2762 at the age of 6 months, demonstrating a 75 kDa product representing the expression of the transgene in the frontal cortex, the temporal cortex, hippocampus, basal ganglia and mesencephalon, but not in the cerebellum or the spinal cord. However, overexposure of the same western blot clearly demonstrates that the transgene is also expressed in the cerebellum and the spinal cord though at a much lower level (data not shown).

Slow progressive phenotypes with emotional, cognitive and motor dysfunction

At birth we found transgenic rats and wild-type littermates phenotypically indistinguishable. Transgenic rats of both sexes are fertile without any sign of atrophy of the sexual organs. We observed a lower body weight gain in transgenic rats that was slowly progressive with the animals being about 20% lighter at the age of 24 months (Fig. 2A). At this age, transgenic rats commonly died after a 2 week period of rapid weight loss, which is associated with emaciation and muscular atrophy (Fig. 2B). Plasma glucose levels were always normal in routine screening (data not shown).

Transgenic animals often showed opisthotonus-like movements of the head. No resting tremor, ataxia, clasping, vocalizations, dyskinesia or seizures were observed. Except for occasional dyskinetic movements of the head, overt behavioral abnormalities were only found on dedicated behavioral testing.

At the age of 2 months transgenic rats developed a reduction of anxiety-like behavior in the elevated plus maze test (Fig. 2C), which is similar to the findings in R6/2 transgenic mice (16). At the age of 10 months transgenic rats showed cognitive decline in a spatial learning task for testing working memory in the radial maze (Fig. 2D and E). At the age of 5 months we had no indication of motor dysfunction in the animals (Fig. 2F), while at the age of 10 and 15 months progressive impairments of hind- and forelimb coordination and balance in the accelerod test were found (Fig. 2G and H). Thus, as in HD patients, emotional and cognitive impairments preceded progressive motor deterioration.

Accumulation of huntingtin aggregates and nuclear inclusions in striatal neurons

We examined whether mutant huntingtin forms aggregates and inclusions in the brain of 18-month-old rats using EM48, a rabbit antibody selective for mutant huntingtin (20,21). Most of the EM48 immunoreactive products appeared as punctuate labeling in the striatum, especially in the ventral region near the lateral ventricles and in the caudal part (Fig. 3B). Occasionally EM48 labeled aggregates were observed in the cortex. Other regions including hippocampus and cerebellum showed very weak or no EM48 label. In wild-type animals no EM48 labeled aggregates or puncta were found (Fig. 3A).

Two types of EM48 labeling, neuropil aggregates and nuclear inclusions were observed. As in other HD animal models (11,22) and in HD patient brains (20) some neuropil aggregates were arranged in linear arrays and most of them were scattered (Fig. 3C). Single nuclear inclusions were mainly observed in the striatum (Fig. 3D), resembling other HD mouse models (14,21). Since the striatal projection neurons terminate their axons in the lateral globus pallidus (LGP), we also examined the caudal region of the striatum. Nuclear staining and neuropil aggregates were common in the striatum. In the LGP, however, most EM48 labeling existed as neuropil aggregates.

To examine at what age mutant huntingtin forms aggregates and inclusions in the ventral region of the striatum, we additionally screened brains of 1-, 6-, 12- and 24-month-old rats for EM48 immunoreactive products (Fig. 3E–H). At the ages of 12 (Fig. 3G), 18 (Fig. 3A–D), and 24 months punctuate labeling was evident, which was most pronounced at the age of 24 months. No aggregates or inclusions were found in the brain of 1- and 6-month-old rats.

Postmortem concentrations of tryptophan and biogenic amines

Since altered tryptophan and dopamine metabolism is linked to HD, we examined neurochemical alterations in the transgenic HD rats using a highly sensitive HPLC method (23). Striatal dopamine levels were decreased only about 20% in heterozygotic rats whereas in homozygotic rats a reduction of nearly 80% was found (Fig. 4A). The levels of dopamine and DOPAC in the parietal cortex of homozygotic animals were not significantly changed (Fig. 4B, D and E). Tryptophan concentrations were decreased 2-fold in striatum (Fig. 4E), but not significantly different in parietal cortex (Fig. 4F). Interestingly, the levels of xanturenic acid were nearly depleted in the striatum of homozygotic transgenic rats (Fig. 4G) and undetectable in the parietal cortex (Fig. 4H). In contrast, in heterozygotes levels of xanturenic acid were elevated in the parietal cortex (Fig. 4H), but unchanged in the striatum (Fig. 4G). No significant changes in other neurotransmitter levels were found.

Focal lesions in the striatum, enlarged lateral ventricles, and reduced brain glucose metabolism

To examine whether transgenic animals display neuropathological signs detectable by magnetic resonance (MR) imaging, we performed MR investigations on 8-month-old homozygotic HD rats. MR scans revealed enlarged lateral ventricles (Fig. 5C and D) and focal lesions in the striatum (Fig. 5F).

Since clinical studies have consistently revealed reductions in striatal glucose metabolism, we studied the local cerebral metabolic rate of glucose ($lCMR_{Glc}$) in transgenic rats using [^{18}F]FDG (fluor-deoxy-glucose) and a high-resolution small-animal PET (positron emission tomography). PET studies were accompanied by *ex vivo* [^{18}F]FDG measurements in order to test their reliability.

Harderian glands and different parts of the brain, such as olfactory bulb and caudato-putamen, were clearly distinguishable (Fig. 6). Individually co-registered MR images allowed a precise delineation of the whole brain as region of interest (ROI), as indicated by the red line (Fig. 6A and E). The defined ROI was measured in the co-registered PET image (Fig. 6B–D, F–H). Mean $lCMR_{Glc}$ values, as calculated from animal PET data of control animals, were 54.98 ± 15.53 [μmol/(100 g × min)] for the whole brain. Mean $lCMR_{Glc}$ values of hetero- and homozygotic animals were lower than control values (see legend of Fig. 6). Metabolic abnormalities of homozygotic animals were significantly different from controls ($P < 0.05$).

After completion of the PET scanning, we subsequently acquired $lCMR_{Glc}$ values *ex vivo* using [^{18}F]FDG autoradiography (Fig. 6J and K). Similar to the *in vivo* situation determined by [^{18}F]FDG-PET, mean *ex vivo* $lCMR_{Glc}$ values of homozygotic animals were significantly lower than control values ($P < 0.05$). A statistical comparison of autoradiographic

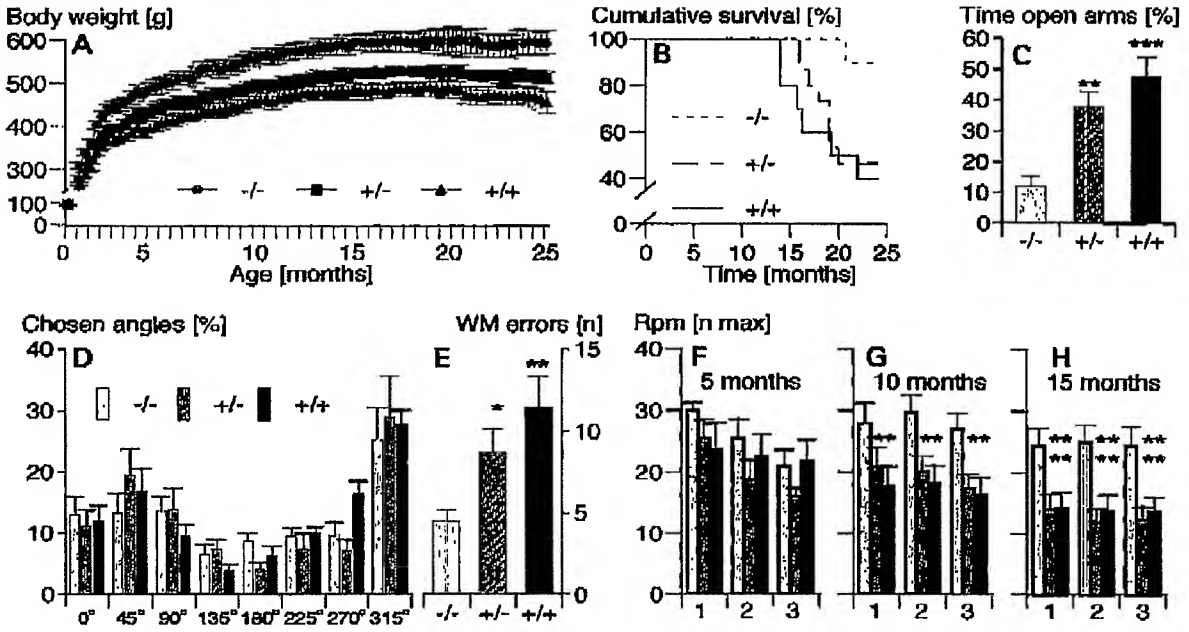


Figure 2. Growth, survival, and behavioral phenotyping. Growth chart representing absolute body weight measured once a week of male wild-type (−/−; gray round symbols) and HD transgenic (+/−, dark squares and +/+, black triangles) rats from 1 to 24 months of age (A). Symbols indicate means± SEM. A significant effect of genotype ($P < 0.001$) and a significant genotype × weight-gain interaction ($P < 0.0001$) indicate a progressive decline in body weight gain in HD transgenics. (B) Cumulative survival of male wild-type (−/−; dotted line) and HD transgenic (+/−, mixed dots/lines and +/+, line) rats from 1 to 24 months of age (end of study) using Kaplan–Meier estimator. Log-rank test revealed a $P < 0.05$. (C) Percentage of time spent on the open arms of the elevated plus maze. Transgenic rats (+/−, hatched columns and +/+, black columns) spent more time (**$P < 0.001$; ***$P < 0.0001$) on the open arms. (D, E) Radial maze behavior. During exploration of the radial maze, transgenic rats showed no major differences in preference for certain angles when choosing arms (D) suggesting that the animals have general motor, cognitive and sensory abilities sufficient to master this task. Activity (total of arm visits and total of time in arms) was not significantly changed (data not shown). Radial maze reinforced alternation demonstrated an increased number of arm visits required to collect all food pellets. The increased number of working memory (WM) errors (E) indicates that the transgene affected the ability to retain and manipulate mnemonic information to guide ongoing behavior (*$P < 0.01$; **$P < 0.001$). Bars indicate means± SEM of each measurement across the trials. (F–H) Balance and motor coordination on the accelerating rod. The means± SEM of the maximal speed (rpm) and the duration of balance (data not shown) were recorded. At the age of 5 months HD transgenic rats were not significantly impaired in their ability to stay on the rotating rod (F). At the age of 10 and 15 months HD transgenic rats exhibit difficulty and a progressive decline in performance on the accelerod (G–H). Asterisks indicate significant differences between wild-type (−/−) control and homo- as well as heterozygotic HD transgenic rats (*$P < 0.01$; **$P < 0.001$).

and animal PET data indicated that $lCMR_{Glc}$ values were significantly similar ($P < 0.05$).

DISCUSSION

In this report we describe the first transgenic rat model for Huntington's disease, which displays symptoms similar to the most frequent late-onset form of HD. It should be emphasized that these transgenic rats represent the first animal model of a human neurodegenerative disorder of the brain *per se* and that these animals express a high-end adult-onset HD allele, which is associated with a slow disease progression and pathology restricted to the striatum. Other symptomatic transgenic mice, however, express very large repeats that are only found in juvenile HD patients. Thus, these HDtg rats are especially useful for studying pathological changes that may be commonly present in the majority of adult HD patients, making this rat model more valuable than other mouse models in evaluating novel therapeutics on HD.

Transgenic rats develop slowly progressive phenotypes with emotional, cognitive, and motor deteriorations. The emotional disturbance is characterized by a reduction of anxiety, which resembles similar observations in R6/2 HD transgenic mice (16). Cognitive decline is also a feature of HD (24). Early in the course of HD, patients frequently show impairments of spatial working memory (25), and comparable deficits are also found in R6/2 mice (15,26) as well as in our HD transgenic rats. These data suggest a common underlying neuropathological mechanism in HD and corresponding animal models.

Neuropathological examination revealed nuclear inclusions and neuropil aggregates. EM48 labeled aggregates are mainly found in the striatum of transgenic rats at the age of one year and older. EM48 labeling shows a distribution pattern similar to that in the human condition (20). Similar results were previously reported in HD knock-in mice expressing full-length

Figure 3. EM48 immunostaining of brains of wild-type and HD transgenic rats. (A, B) Low magnification of micrographs of wild-type (A) and HD (B) rat brains. Note that EM48 immunoreactive product is particularly enriched in the ventral part of striatum (Str) near the ventricle (arrow) in HD rat brain. Ctx, cortex. Scale bar, 50 µm. (C) In the caudate part of the striatum of HD rats, many nuclear aggregates and small neuropil aggregates are evident. Neuropil aggregates (arrows) are also present in the lateral globus pallidus (LGP). Scale bar, 25 µm. (D) High magnification of micrograph showing that both EM48 labeled nuclear inclusion (arrowheads) and small neuropil aggregates (arrows) are present in the striatum of HD rat brain. (E–H) Corresponding micrographs of coronal sections at the level of the bregma of 1-month (E), 6-month (F), 12-month (G), 24-month old transgenic rats. Scale bar, 10 µm.

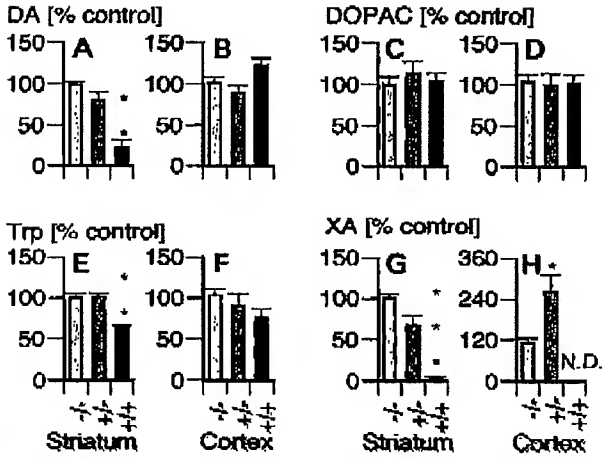


Figure 4. Regional alterations in tryptophan metabolism in HD transgenic rats. The levels of dopamine (A, B), DOPAC (C, D), tryptophan (E, F) and xanturenic acid (G, H) in striatum (A, C, E, G) or parietal cortex (B, D, F, H) in wild-type (−/−) or homo-(+/+) and heterozygotic (+/−) transgenic rats expressing human HD mutation at the age of 18 months. Asterisks indicate significant differences from control rats (*$P < 0.05$. **$P < 0.001$, ***$P < 0.0001$).

mutant huntingtin under the endogenous mouse HD promoter (21,27). A remarkable observation in neurochemistry was that xanturenic acid was nearly completely depleted in the striatum and the parietal cortex. The levels of xanturenic acid were higher in the less afflicted heterozygotes, perhaps reflecting a neurochemical defense mechanism against the excitotoxicity of the overactive indoleamine (2,3)-dioxygenase pathway (28). Similar to HD patients, the levels of tryptophan were decreased in the striatum of homozygotes. Decreased DA and normal DOPAC levels are indicative of increased DA turnover. Decreased levels of tryptophan may be related to an increased formation of quinolinic acid, a neuroexitant molecule with neurotoxic properties (5). These findings support the hypothesis that both increased formation of quinolinic acid (28) and decreased production of neuroprotective metabolites from tryptophan (29) may be relevant to the pathogenesis of HD.

An important feature of the presented HD rat model is its suitability for *in vivo* metabolic and structural imaging, which cannot yet be achieved with transgenic mice. MR scanning demonstrated an enlargement of the lateral ventricles and focal signal abnormalities in the striatum of HD transgenic animals, although quantitative assessment of striatal neurons revealed no significant cell loss. This indicates that striatal atrophy depicted by MR imaging is rather a consequence of shrinkage than neuronal death. In high-resolution animal PET we found a significant reduction of brain glucose metabolism in 2-year-old homozygotic HD rats. In late stages of human HD, clinical PET studies consistently revealed reduced $ICMR_{Glc}$ in the striatum (30,31). Thus, this report provides evidence that the novel HD transgenic rat model does closely resemble the human pathological condition. It is suited for non-invasive *in vivo* investigations of brain metabolism and most probably of further *in vivo* parameters (e.g. receptor density, enzyme activity). Brain atrophy and extracranial tracer accumulation, however, necessitate the application of high-resolution tomographs and a careful evaluation of partial volume and spill over effects.

We report the successful development of a transgenic rat model of HD, which expresses a high-end adult onset HD allele with 51 CAG repeats and which exhibits a high degree of similarity to the most frequent adult type of the disease, thereby permitting *in vivo* monitoring of individual disease progression by high-resolution imaging (PET and MRI). For the first time it is now possible to follow up disease progression in longitudinal *in vivo* studies and to monitor the effects of long-term treatments, microsurgery, neuronal cell transplantation, or antisense approaches on discourse of experimental HD.

MATERIALS AND METHODS

Generation of transgenic rats

To generate the transgene construct, PCR was performed using DNA from a HD patient (19/51 CAGs) with Primer Hu 4

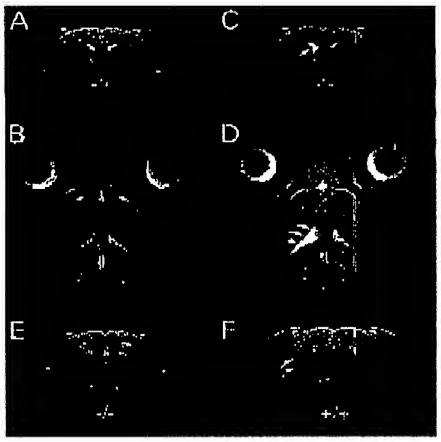


Figure 5. MR scanning of brains of wild-type and HD transgenic rats. (A–D) MR scans of lateral ventricles in coronal (A) and horizontal (B) projection of wild-type (A, B) and HD (C, D) rat brain at the age of 8 months. MR scans of the striatum of a wild-type (E) and an HD transgenic (F) rat brain. Note the enlargement of the lateral ventricles (arrows) and the focal lesions in the striatum (arrows).

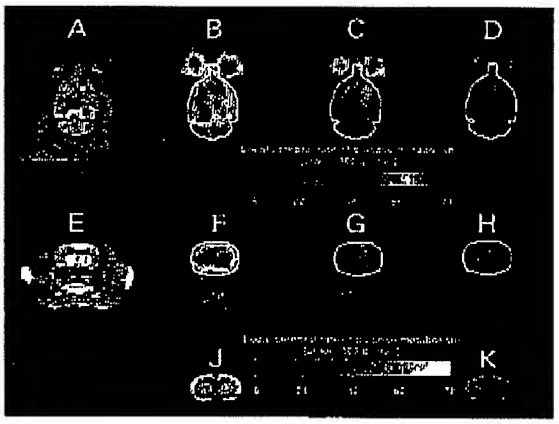


Figure 6. Studies with [^{18}F]FDG and high-resolution small-animal PET. Representative images with [^{18}F]FDG and high-resolution small-animal PET in horizontal (B–D) and coronal (F–H) planes along with individual MR images (A, E) and *ex vivo* autoradiographs (J, K). Individual MR images (A, E) of a control animal are co-registered with respective [^{18}F]FDG-PET images (B, F). Planes are cutting the caudato-putamen level of the brain. Representative sections of *ex vivo* autoradiography (J, K) are taken from identical animals as in [^{18}F]FDG-PET (B, F; D, H). The rat brain is defined within the [^{18}F]FDG-PET on the basis of individually co-registered MR images, as indicated by the red line. Local cerebral rates of glucose metabolism ($lCMR_{Glu}$) are absolutely quantified (see color and black/white bars). The high accumulation of activity in caudato-putamen is clearly visible in [^{18}F]FDG-PET (F, G, H) and *ex vivo* autoradiography (J, K). Homozygotic animals exhibit significantly ($P < 0.05$) lower $lCMR_{Glu}$ values compared with controls, both in [^{18}F]FDG-PET [34.54 ± 18.52 μmol/(100 g × min) versus 54.98 ± 15.53 μmol/(100 g × min)] and in *ex vivo* autoradiography [43.54 ± 6.77 μmol/(100 g × min) versus 63.02 ± 8.24 μmol/(100 g × min)].

(ATGGCGACCCTGGAAAAGCTGATGAA) and Hu3-510 (GGGCGCCTGAGGCTGAGGCAGC). This PCR product was subsequently digested with *Eco81I*. The first 154 nucleotides of the cDNA RHD10 containing nt 1–1962 of the rat HD-gene (18) were removed by restriction of the clone with *Eco*RV and *Eco*81I. This fragment was replaced by the PCR product. Subsequently, a 885 bp rat HD promotor fragment from position −900 to −15 (19) was ligated upstream of the cDNA and a 200 bp fragment containing the SV40 polyadenylation signal was added downstream of the cDNA resulting in RHD/Prom51A. The insert was excised with *Xba*I and *Ssp*I out of the cloning vector and microinjected into oocyte donors of Sprague–Dawley (SD) rats (32,33). Tail DNA was extracted from each of the offspring and Southern blots of *EcoRI* digested DNA were performed to screen for founders.

For western blot analysis, frozen brain halves and dissected brain areas were homogenized and protein extracted. Protein extracts were subjected to SDS–PAGE and blotted electrophoretically onto Immobilon-P membranes. Detection of huntingtin protein was performed basically as described (34) using the polyclonal anti-huntingtin antibody 675.

Behavioral phenotyping of the HD transgenic rat line

The considerations for behavioral phenotyping of transgenic and knockout mice (35) were adapted with specific modifications for testing rats. All procedures were approved by the Government of Lower Saxony in Hannover, Germany, and performed in compliance with international animal welfare standards. The elevated plus maze (TSE-Systems, Bad Homburg, Germany) was equipped with light beam sensors and had two open arms (50 × 10 cm) and two enclosed arms of the same size. The experiment was conducted with 2-month-old rats as previously described (36). An increase of the time spent in the open arms is interpreted as an anxiolytic-like response. An automated sensor-equipped eight-arm radial maze (TSE) was used to measure learning and memory in an experimental design testing exploring behavior and working memory (WM) errors in allocentric orientation (37). An accelerating rotarod for rats (TSE) was used to measure fore- and hind-limb motor coordination and balance. Training consisted of three trials per day on four consecutive days. The duration of each trial was 5 min on accelerating mode of the apparatus. The maximal speed level and the mean latency to fall off the rotarod were recorded on three consecutive tests. Data were subjected to one- or two-way ANOVA with one between-subject factor (genotype) and with repeated measurements on one or more factors depending on the test used. The PLSD test was used for *post hoc* comparisons. Cumulative survival was calculated by means of Kaplan–Meier analysis. A critical value for significance of $P < 0.05$ was used throughout the study.

Immunohistology and light microscopic examination

Brains of HD transgenic rats and controls at the age of 1, 6, 12, 18 and 24 months were perfused intracardially with PBS followed by paraformaldehyde and postfixed. Free-floating sections were pre-blocked in normal goat serum, Triton-X and

avidin, and incubated with EM48 antibody (1:400 dilution) at 4°C for 24 h (20,21). The EM48 immunoreactive product was visualized with the avidin–biotin complex kit (Vector ABC Elite, Burlingame, CA, USA).

Analysis of neurotransmitters from post-mortem tissue samples

Tryptophan and its kynurenine, catechol- and indoleamine metabolites were measured by electrochemical HPLC, as described previously (23). Briefly, striatum and parietal cortex of 18-month-old transgenic HD rats were dissected, weighed and sonicated in perchloric acid. The homogenate was centrifuged and 20 µl of supernatant was injected into a HPLC system (ESA model 5600 CoulArray module, Chelmsford, MA, USA) with two coulometric arraycell modules, each with four working electrodes. The chromatographic separation was achieved on an ESA MD-150 reversed-phase C_{18} analytical column with a Hypersil pre-column.

MR scanning

Rats were anesthetized with 2% isoflurane and fixed in a stereotaxic frame. MRI was performed on a 4.7 T Bruker Biospec scanner with a free-bore of 20 cm equipped with an actively RF-decoupled coil system. A whole-body birdcage resonator enabled homogeneous excitation, and a 3 cm surface coil was used as receiver. T_2-weighted spin echo images were acquired using a rapid acquisition relaxation enhanced (RARE) sequence (38). Eleven axial and seven coronal slices were measured (slice thickness: 1.5 mm axial; 1.3 mm coronal; field of view, 3.2 × 3.2 cm; matrix, 256 × 256; *TR*/*TE* 3000/19 ms; six averages).

PET studies

PET imaging was performed on a dedicated high-resolution small-animal PET scanner ('TierPET') as previously described (39) on 24-month-old homozygotic (+/+; $n=6$) and heterozygotic animals (+/−; $n=7$), as well as age-matched controls (−/−; $n=6$). Reconstructed image resolution was 2.1 mm, which is homogeneously maintained throughout the entire field of view. A precise anatomical identification of rat brain regions was achieved by co-registration of magnetic resonance (MRI; Siemens Magnetom, 1.5 T, equipped with a dedicated small limb coil) and PET images. Animals received an injection of 0.3 ml [^{18}F]FDG (1 mCi/ml, solved in NaCl 0.9%) under isoflurane sedation. After 30 min animals were anesthetized with ketamine/xylazine and glucose concentrations and input function were detected by serial blood samples. After a 60 min PET scan brains were removed and immediately frozen. Cryostat sections (20 µm) were exposed to a phosphor imaging plate (BAS-SR 2025, Fuji, Germany) together with calibrated fluorine-18 brain paste standards. Imaging plates were scanned with a high-performance imaging plate reader (BAS5000 BioImageAnalyzer, Fuji, Germany; spatial resolution, 50 µm). Local cerebral metabolic rate of glucose ($ICMR_{Glc}$) was calculated on the basis of the operational equation used in 2DG autoradiography studies (40) with modified rate and lumped constants to account for the difference in kinetic characteristics between FDG and 2DG. The following constants (41) were used: $k_1=0.30$; $k_2=0.40$; $k_3=0.068$; lumped constant, $LC=0.60$. Similarity of $ICMR_{Glc}$ as determined by FDG-PET and *ex vivo* autoradiography was analyzed by linear regression analysis.

ACKNOWLEDGEMENTS

We thank R. Barnow for excellent technical assistance and S. Fryk for the correction of the English. We gratefully acknowledge S. Weber, M. Cremer, U. Pietrzyk, and T. Rustige for their technical assistance with the *TierPET* and N.J. Shah for his help with high-resolution MRI (all Research Center Jülich). [^{18}F]FDG was provided in clinical purity by the Institute of Nuclear Chemistry (Research Center Jülich). This work was supported by a Volkswagen Foundation grant to R.P. and S.v.H. (I/75169), by NIH grants NS41669 and AG19206 to X.J.L., by HGF strategy funds to A.B. (project 2000/09: 'Primate PET'), and by a BMBF grant to O.R. (01KV9523/6).

REFERENCES

1. The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
2. Harper, P.S. (1996) *Huntington's Disease*. W.B. Saunders, London.
3. Borlongan, C.V., Koutouzis, T.K., Freeman, T.B., Cahill, D.W. and Sanberg, P.R. (1995) Behavioral pathology induced by repeated systemic injections of 3-nitropropionic acid mimics the motoric symptoms of Huntington's disease. *Brain Res.*, **697**, 254–257.
4. Brouillet, E., Hantraye, P., Ferrante, R.J., Dolan, R., Leroy-Willig, A., Kowall, N.W. and Beal, M.F. (1995) Chronic mitochondrial energy impairment produces selective striatal degeneration and abnormal choreiform movements in primates. *Proc. Natl Acad. Sci. USA*, **92**, 7105–7109.
5. Miranda, A.F., Boegman, R.J., Beninger, R.J. and Jhamandas, K. (1997) Protection against quinolinic acid-mediated excitotoxicity in nigrostriatal dopaminergic neurons by endogenous kynurenic acid. *Neuroscience*, **78**, 967–975.
6. Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S.W. *et al.* (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, **87**, 493–506.
7. Reddy, P.H., Williams, M., Charles, V., Garrett, L., Pike-Buchanan, L., Whetsell, W.O., Jr., Miller, G. and Tagle, D.A. (1998) Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. *Nat. Genet.*, **20**, 198–202.
8. Hodgson, J.G., Agopyan, N., Gutekunst, C.A., Leavitt, B.R., LePiane, F., Singaraja, R., Smith, D.J., Bissada, N., McCutcheon, K., Nasir, J. *et al.* (1999) A YAC mouse model for Huntington's disease with full-length mutant huntingtin, cytoplasmic toxicity, and selective striatal neurodegeneration. *Neuron*, **23**, 181–192.
9. Schilling, G., Becher, M.W., Sharp, A.H., Jinnah, H.A., Duan, K., Kotzuk, J.A., Slunt, H.H., Ratovitski, T., Cooper, J.K., Jenkins, N.A. *et al.* (1999) Intranuclear inclusions and neuritic aggregates in transgenic mice expressing a mutant N-terminal fragment of huntingtin. *Hum. Mol. Genet.*, **8**, 397–407.
10. Shelbourne, P.F., Killeen, N., Hevner, R.F., Johnston, H.M., Tecott, L., Lewandoski, M., Ennis, M., Ramirez, L., Li, Z., Iannicola, C. *et al.* (1999) A Huntington's disease CAG expansion at the murine Hdh locus is unstable and associated with behavioural abnormalities in mice. *Hum. Mol. Genet.*, **8**, 763–774.
11. Yamamoto, A., Lucas, J.J. and Hen, R. (2000) Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell*, **101**, 57–66.

12. Dunnett, S.B., Carter, R.J., Watts, C., Torres, E.M., Mahal, A., Mangiarini, L., Bates, G. and Morton, A.J. (1998) Striatal transplantation in a transgenic mouse model of Huntington's disease. *Exp. Neurol.*, **154**, 31–40.
13. Carter, R.J., Lione, L.A., Humby, T., Mangiarini, L., Mahal, A., Bates, G.P., Dunnett, S.B. and Morton, A.J. (1999) Characterization of progressive motor deficits in mice transgenic for the human Huntington's disease mutation. *J. Neurosci.*, **19**, 3248–3257.
14. Davies, S.W., Turmaine, M., Cozens, B.A., DiFiglia, M., Sharp, A.H., Ross, C.A., Scherzinger, E., Wanker, E.E., Mangiarini, L. and Bates, G.P. (1997) Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell*, **90**, 537–548.
15. Lione, L.A., Carter, R.J., Hunt, M.J., Bates, G.P., Morton, A.J. and Dunnett, S.B. (1999) Selective discrimination learning impairments in mice expressing the human Huntington's disease mutation. *J. Neurosci.*, **19**, 10428–10437.
16. File, S.E., Mahal, A., Mangiarini, L. and Bates, G.P. (1998) Striking changes in anxiety in Huntington's disease transgenic mice. *Brain Res.*, **805**, 234–240.
17. Jacobs, R.E. and Cherry, S.R. (2001) Complementary emerging techniques: high-resolution PET and MRI. *Curr. Opin. Neurobiol.*, **11**, 621–629.
18. Schmitt, I., Bechner, D., Megow, D., Henklein, P., Hameister, H., Epplen, J.T. and Riess, O. (1995) Expression of the Huntington disease gene in rodents: cloning the rat homologue and evidence for downregulation in non-neuronal tissues during development. *Hum. Mol. Genet.*, **4**, 1173–1182.
19. Holzmann, C., Maueler, W., Petersohn, D., Schmidt, T., Thiel, G., Epplen, J.T. and Riess, O. (1998) Isolation and characterization of the rat huntingtin promoter. *Biochem. J.*, **336**, 227–234.
20. Gutekunst, C.A., Li, S.H., Yi, H., Mulroy, J.S., Kuemmerle, S., Jones, R., Rye, D., Ferrante, R.J., Hersch, S.M. and Li, X.J. (1999) Nuclear and neuropil aggregates in Huntington's disease: relationship to neuropathology. *J. Neurosci.*, **19**, 2522–2534.
21. Li, H., Li, S.H., Johnston, H., Shelbourne, P.F. and Li, X.J. (2000) Amino-terminal fragments of mutant huntingtin show selective accumulation in striatal neurons and synaptic toxicity. *Nat. Genet.*, **25**, 385–389.
22. Li, H., Li, S.H., Cheng, A.L., Mangiarini, L., Bates, G.P. and Li, X.J. (1999) Ultrastructural localization and progressive formation of neuropil aggregates in Huntington's disease transgenic mice. *Hum. Mol. Genet.*, **8**, 1227–1236.
23. Vaarmann, A., Kask, A. and Maeorg, U. (2002) Novel and sensitive high-performance liquid chromatographic method based on electrochemical coulometric array detection for simultaneous determination of catechol-amines, kynurenine and indole derivatives of tryptophan. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **769**, 145–153.
24. Mohr, E., Brouwers, P., Claus, J.J., Mann, U.M., Fedio, P. and Chase, T.N. (1991) Visuospatial cognition in Huntington's disease. *Mov. Disord.*, **6**, 127–132.
25. Lawrence, A.D., Sahakian, B.J., Hodges, J.R., Rosser, A.E., Lange, K.W. and Robbins, T.W. (1996) Executive and mnemonic functions in early Huntington's disease. *Brain*, **119**, 1633–1645.
26. Murphy, K.P., Carter, R.J., Lione, L.A., Mangiarini, L., Mahal, A., Bates, G.P., Dunnett, S.B. and Morton, A.J. (2000) Abnormal synaptic plasticity and impaired spatial cognition in mice transgenic for exon 1 of the human Huntington's disease mutation. *J. Neurosci.*, **20**, 5115–5123.
27. Wheeler, V.C., White, J.K., Gutekunst, C.A., Vrbanac, V., Weaver, M., Li, X.J., Li, S.H., Yi, H., Vonsattel, J.P., Gusella, J.F. *et al.* (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.*, **9**, 503–513.
28. Bruyn, R.P. and Stoof, J.C. (1990) The quinolinic acid hypothesis in Huntington's chorea. *J. Neurol. Sci.*, **95**, 29–38.
29. Stone, T.W. (2001) Kynurenines in the CNS: from endogenous obscurity to therapeutic importance. *Prog. Neurobiol.*, **64**, 185–218.
30. Kuwert, T., Lange, H.W., Langen, K.J., Herzog, H., Aulich, A. and Feinendegen, L.E. (1990) Cortical and subcortical glucose consumption measured by PET in patients with Huntington's disease. *Brain*, **113**, 1405–1423.
31. Young, A.B., Penney, J.B., Starosta-Rubinstein, S., Markel, D.S., Berent, S., Giordani, B., Ehrenkaufer, R., Jewett, D. and Hichwa, R. (1986) PET scan investigations of Huntington's disease: cerebral metabolic correlates of neurological features and functional decline. *Ann. Neurol.*, **20**, 296–303.
32. Mullins, J.J., Peters, J. and Ganten, D. (1990) Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene. *Nature*, **344**, 541–544.
33. Schinke, M., Baltatu, O., Bohm, M., Peters, J., Rascher, W., Bricca, G., Lippoldt, A., Ganten, D. and Bader, M. (1999) Blood pressure reduction and diabetes insipidus in transgenic rats deficient in brain angiotensinogen. *Proc. Natl Acad. Sci. USA*, **96**, 3975–3980.
34. Schmidt, T., Landwehrmeyer, G.B., Schmitt, I., Trottier, Y., Auburger, G., Laccone, F., Klockgether, T., Volpel, M., Epplen, J.T., Schols, L. *et al.* (1998) An isoform of ataxin-3 accumulates in the nucleus of neuronal cells in affected brain regions of SCA3 patients. *Brain Pathol.*, **8**, 669–679.
35. Crawley, J.N. (1999) Behavioral phenotyping of transgenic and knockout mice: experimental design and evaluation of general health, sensory functions, motor abilities, and specific behavioral tests. *Brain Res.*, **835**, 18–26.
36. Breivik, T., Stephan, M., Brabant, G.E., Straub, R.H., Pabst, R. and von Horsten, S. (2002) Postnatal lipopolysaccharide-induced illness predisposes to periodontal disease in adulthood. *Brain Behav. Immun.*, **16**, 421–438.
37. Holscher, C. and Schmidt, W.J. (1994) Quinolinic acid lesion of the rat entorhinal cortex pars medialis produces selective amnesia in allocentric working memory (WM), but not in egocentric WM. *Behav. Brain Res.*, **63**, 187–194.
38. Hennig, J., Nauerth, A. and Friedburg, H. (1986) RARE imaging: a fast imaging method for clinical MR. *Magn. Reson. Med.*, **3**, 823–833.
39. Weber, S., Bauer, A., Herzog, F., Kehren, H., Mühlensiepen, J., Vogelbruch, H., Coenen, H., Zilles, K. and Halling, H. (2000) Recent results of the TierPET scanner. *IEEE Trans. Nucl. Sci.*, **47**, 1665–1669.
40. Sokoloff, L., Reivich, M., Kennedy, C., Des Rosiers, M.H., Patlak, C.S., Pettigrew, K.D., Sakurada, O. and Shinohara, M. (1977) The [14C]deoxy-glucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat. *J. Neurochem.*, **28**, 897–916.
41. Ackermann, R.F. and Lear, J.L. (1989) Glycolysis-induced discordance between glucose metabolic rates measured with radiolabeled fluorodeoxyglucose and glucose. *J. Cereb. Blood Flow. Metab.*, **9**, 774–785.

EXHIBIT 8

 Human Molecular Genetics, 1997, Vol. 6, No. 10 Review 1633–1637

Transgenic models of Huntington's disease

Gillian P. Bates*, Laura Manglarini, Amarbirpal Mahal and Stephen W. Davies[1]

Medical and Molecular Genetics, UMDS, Guy's Hospital, London SE1 9RT, UK and [1]Department of Anatomy and Developmental Biology, University College London, Gower Street, London WC1E 6BT, UK

Received May 7, 1997

CAG/polyglutamine expansion has been shown to form the molecular basis of an increasing number of inherited neurodegenerative diseases. The mutation is likely to act by a dominant gain of function but the mechanism by which it leads to neuronal dysfunction and cell death is unknown. The proteins harbouring these polyglutamine tracts are unrelated and without exception are widely expressed with extensively overlapping expression patterns. The factors governing the cell specific nature of the neurodegeneration have yet to be understood. Upon a certain size threshold, expanded CAG repeats become unstable on transmission and a modest degree of somatic mosaicism is apparent. Similarly, the molecular basis of the instability and its tissue specificity has yet to be unravelled. Recent reports describing the first mouse models of CAG/polyglutamine disorders indicate that it will be possible to model both the pathogenic mechanism and the CAG repeat instability in the mouse. This has great potential and promise for uncovering the molecular basis of these diseases and developing therapeutic interventions.

INTRODUCTION

Huntington's disease (HD) (1) is one of an increasing number of neurodegenerative disorders caused by a CAG/polyglutamine (polygln) repeat expansion, including spinal and bulbar muscular atrophy (SBMA) (2), dentatorubral pallidoluysian atrophy (DRPLA) (3,4) and spinocerebellar ataxia (SCA) types 1 (5), 2 (6–8), 3 (9) and 6 (10). The inheritance patterns are autosomal dominant (with the exception of X-linked SBMA) and in each case, the proteins can tolerate a large variation in the size of the polygln tracts in the normal range but upon a certain size (~37–40 glutamines) these tracts become pathogenic. It is likely that the novel molecular pathways initiated by this mutation have a common basis (except possibly in the case of SCA6 in which the pathogenic threshold is smaller). The proteins harbouring the polygln stretches are mostly novel and otherwise unrelated. In all cases the proteins are widely or ubiquitously expressed, but despite extensively overlapping expression patterns, the neuronal cell death is relatively specific and can differ markedly (reviewed in 11).

The molecular events by which a polygln expansion causes cell death remain to be unravelled (reviewed in 12). These mutations are likely to act by a dominant gain of function, this mechanism being supported by the identification of the 1C2 antibody which specifically detects polygln expansions, suggestive of a conformational change at a certain size threshold (13). In addition, the factors which convey the specific and differing patterns of cell death between these diseases are not understood. Possible mechanisms include differences in expression levels, subcellular localisation of the mutated protein or cell specific subcellular interactions. A number of proteins have now been reported to interact with huntingtin which include HAP 1 (14), HIP-1 (15), a specific ubiquitin-conjugating enzyme (16) and GAPDH (17). It is yet to be established whether any of these proteins play a role in the pathogenic mechanism. Huntingtin has also been shown to be specifically cleaved by apopain, a cysteine protease with a key role in the proteolytic events leading to apoptosis (18). Similarly, it is not clear if this participates in the chain of events leading to neurodegeneration.

Expanded triplet repeats are invariably unstable when inherited from one generation to the next and they generally show varying degrees of somatic mosaicism. The intergenerational instability forms the molecular basis of anticipation: the observation that the age of onset of a disease decreases and/or the severity increases as the gene is passed from one generation to the next. Repeat instability on transmission has been described in all of the CAG repeat diseases and, in general, repeats tend to be more unstable on paternal transmission. This may present as larger increases on paternal inheritance as in HD (19) (reflected in the paternal sex bias to the anticipation) or as a tendency to increase on male and decrease on female transmission as in SCA1 (20). A relatively modest degree of somatic repeat instability has been identified in HD, DRPLA, SCA1 and MJD. In general, expansions have been identified in regions of the CNS, with the exception of the cerebellum which presents a smaller repeat relative to the other brain regions tested (21–26). Of non-CNS tissues, instability has consistently been reported in liver and kidney (21,24–26) and also in muscle, lung, testis (21), leukocytes (23) and colon (24,26). Studies of DRPLA patients also identified a significant correlation between the range of the expanded allele and the age at death of the patient rather than with the onset of disease (25). The molecular events governing triplet repeat instability are not understood and possible mechanisms must address both a CAG repeat size threshold and cell specificity.

TRANSGENIC MODELLING OF HUNTINGTON'S DISEASE

It has been proposed for many years that the HD mutation most probably acts through a dominant gain of function. Analysis of mice arising from the first transgenic models of HD, SCA1 and

*To whom correspondence should be addressed. Tel: +44 171 955 4485; Fax: +44 171 95 4444; Email: g.bates@umds.ac.uk

MJD in addition to gene targeted knockouts of the mouse HD gene (*Hdh*) supports this hypothesis.

Knockouts of the mouse *Hhd* gene

Three research groups have independently generated knockouts of the mouse HD gene (*Hdh*) (27–29). In all cases the nullizygous phenotype was embryonic lethal, clearly demonstrating that the HD gene plays an important role in development. In two of these studies, heterozygous mice expressing only one copy of *Hdh* were phenotypically normal (28,29). In contrast, Nasir *et al.* (27) reported that heterozygotes showed increased motor activity and cognitive deficits with a significant neuronal loss in the subthalamic nucleus. To explain this discrepancy, it has been suggested that the targeted allele (targeted to replace exon 5) may allow the production of a truncated protein which could conceivably cause a dominant effect in the heterozygous mice, generating a phenotype. Together, these studies demonstrate that HD is not caused by haplo-insufficiency (loss of function of one copy of the gene) or a simple dominant-negative mechanism. In the first case, loss of one allele and in the second case, loss of both alleles, would be expected to generate a model of HD.

Transgenic models of HD

A dominant gain of function mechanism would predict that a mouse model of a polyGln neurodegenerative disorder could be generated by the introduction into the mouse germ line of a mutant copy of the gene in question, irrespective of the presence of two copies of the endogenous mouse homologue. The first description of HD transgenic mice used a full length cDNA construct under the control of a CMV promoter carrying $(CAG)_{44}$ (30). Of the HD transgenes, 2/6 founders expressed high levels of transgene mRNA but a transgene protein was not detected. Whilst these results could be interpreted as providing evidence that translation of the CAG repeat into a polyGln expansion is necessary for pathogenesis, the repeat expansion in this experiment is comparatively modest and it is possible that larger expansions are necessary to generate a phenotype with an age of onset that falls within the lifetime of a mouse.

Genomic clones are frequently more successful at generating transgenic models than cDNAs as they often direct an expression profile that mimics the endogenous gene. The large size of the HD gene (170 kb) necessitates that genomic constructs are prepared and manipulated in the form of yeast artificial chromosomes (YACs). Using YAC technology, Hodgson *et al.* (31) have successfully generated mice that are transgenic for the normal human HD gene. They have crossed the human HD transgene onto an *Hdh* nullizygous background and shown that the human YAC can rescue the embryonic lethal phenotype. This indicates that the transgene is expressed appropriately and predicts that the introduction of a mutant version of the human YAC would be successful in generating a model of HD.

Mice transgenic for a mutant version of exon 1 of the HD gene

We have described four lines of mice that are transgenic for exon 1 of the HD gene carrying CAG expansions of 115–156 (R6/1, R6/2, R6/5 and R6/0) and a further two lines transgenic for the same construct carrying 18 repeats (HDex6 and HDex27) (32). The transgene is ubiquitously expressed at both the RNA and protein levels in all lines except R6/0, in which no evidence of expression has been detected (32,33). The transgene protein contains the first 69 amino acids of huntingtin in addition to the number of residues encoded by the CAG repeat (i.e. ~3% of huntingtin).

A progressive neurological phenotype has been observed in three lines: R6/1, $(CAG)_{115}$; R6/2, $(CAG)_{145}$; and R6/5, $(CAG)_{130–155}$. Line R6/2 has an onset of ~2 months, line R6/1 at ~5 months and R6/5 hemizygotes do not show symptoms after >1 year. Lines R6/1 and R6/5 show an earlier age of onset and more rapid progression of the disease when bred to homozygosity. The phenotype includes an irregular gait, resting tremor, stereotypic and abrupt, irregularly timed movements and epileptic seizures. Coincident with the onset of the motor disorder there is a progressive reduction in body weight in the transgenes as compared with their littermate controls. The absence of a phenotype in lines R6/0, HDex6 and HDex27 suggests that expression of the polyGln expansion forms the molecular basis of the phenotype rather than the expression of a novel peptide. It is notable that the R6 phenotype does not include an overt cerebellar ataxia as described for the spinocerebellar ataxia lines (34,35) (see below). Extensive neuropathological analysis has been performed on the brains of R6/2 mice. At 12 weeks, the only difference that could be identified between the transgene and control brains was that the R6/2 brains were ~20% smaller, and that this reduction in brain size occurred across all structures with an apparently normal neuronal density. This is consistent with early changes that occur in the brains of HD patients. More recently, immunocytochemistry with huntingtin N-terminal antibodies has identified the presence of neuronal intranuclear inclusions (NII) in the brains of symptomatic transgenic mice (33).

COMPARISON WITH OTHER CAG/POLYGLN MOUSE MODELS

Mice transgenic for both SCA1 (34) and MJD (35) constructs have also been reported to develop a phenotype. A summary of the main features of these and the R6 transgenes is presented in Table 1. The SCA1 transgenes were the first demonstration that modelling a polyGln repeat disorder would be possible in the mouse. They included mice transgenic for the SCA1 cDNA carrying either a normal interrupted allele of $(CAG)_{12}CATCAGCAT(CAG)_{15}$ (PS-30) or an expanded uninterrupted allele of $(CAG)_{82}$ (PS-82) under the control of the pcp2 promoter (Purkinje cell-specific) (34). Five of six PS-82 lines showed RNA expression between 10- and 100-fold of endogenous levels. In the original report, transgene protein could not be detected but this has since been shown to be present by immunocytochemistry (H.Orr and H.Zoghbi, personal communication). Mice from all five lines developed ataxia. Onset varied from 12 to 26 weeks and a dosage effect was apparent: in two lines studied, homozygotes were more severely affected than hemizygotes. Neuropathological analysis showed significant loss of the Purkinje cell population, with Bergmann glial proliferation, and shrinkage and gliosis of the molecular layer. Ectopic Purkinje cells were present in the molecular layer and occasionally the granular layer and the dendritic arrays also appeared to be abnormal.

Ikeda *et al.* (35) used expression constructs containing the MJD cDNA carrying 79 CAG repeats (MJD79), the CAG repeat followed by only the C-terminus of the MJD gene with both 79 and 35 CAG repeats ($Q_{79}C$ and $Q_{35}C$) and a 79 CAG repeat in

isolation (Q_{79}) under the control of a Purkinje cell-specific promoter. Ataxia was observed in 3/3 $Q_{79}C$ and 2/6 Q_{79} transgenic mice, occurring as early as 1 month of age after full activation of the promoter. In contrast, a phenotype was not observed in any of the ten $Q_{35}C$ or four MJD79 transgenic mice as of 7 and 5 months, respectively. Neuropathological analysis of a 2 month old ataxic Q79C mouse showed a very atrophic cerebellum, in which all three layers were affected. The authors suggest that comparison of their data with that of Burright *et al.* (34) indicates that the polyGln tracts are more toxic when present in isolation or in the context of a truncated protein. In the absence of any data relating to expression levels it is difficult to come to strong conclusions with regard to the comparative toxicity of the full length and truncated constructs. However, these conclusions were strongly supported by a series of transient transfections of COS cells described in the same paper (35).

MOUSE MODELS OF CAG/CTG REPEAT STABILITY

Repeat stability studies carried out on the first mice transgenic for CAG repeat expansions showed no evidence of instability and suggested that the molecular mechanism underlying triplet repeat instability in humans may not exist in the mouse. These initial studies included $(CAG)_{45}$ in the androgen receptor cDNA (36), $(CAG)_{44}$ in the HD cDNA (30), $(CAG)_{82}$ in the SCA1 cDNA (34) and $(CAG)_{79}$ in constructs based on the MJD/SCA3 cDNA (35).

Triplet repeat instability in lines transgenic for the HD mutation (R6)

The R6 lines transgenic for exon 1 of the HD gene carrying $(CAG)_{115}$–$(CAG)_{155}$ expansions showed both intergenerational and somatic repeat instability (32,37). The repeats were clearly unstable on transmission in lines R6/1, R6/2 and R6/5, although this was less clear in line R6/0 as the changes observed in this line could be accounted for by errors in sizing. In line R6/2, the degree of instability increases with the age of the transmitting male (as R6/2 females are sterile it was not possible to look for an age effect on female transmission). R6/5 was the only line in which an extensive comparison of instability on both male and female transmission was conducted and the repeats had a tendency to increase on male transmission and decrease on female transmission (37). This trend was supported by the intergenerational instability observed in the other lines. The CAG expansions introduced into these mice are considerably larger than are normally seen in HD patients. The change in size of the repeat on transmission in the mice is smaller than would be expected from comparison with size changes associated with highly expanded CAG repeats seen in humans. The discrepancy in the degree of instability between humans and mice may reflect the difference in their life span, a model supported by the observation that the size of the intergenerational expansion increased with the age of the transmitting male.

Somatic instability was detected in lines R6/1, R6/2 and R6/5 but not in line R6/0 (Fig. 1). In all three lines, onset of instability was at ~6 weeks and the CAG repeat range increased with the age of the mouse. This argues against a pathogenic role for repeat instability as the age of onset of symptoms in these lines differs markedly. The pattern of instability was more widespread in some lines than others although on the whole it was first present and most prominent in brain regions. Peripheral tissues that consistently showed instability included liver and kidney. Overall the somatic instability was comparable with that described in individuals carrying CAG expansions (21–26). The major difference between line R6/0, in which instability was not apparent, and the other lines was the absence of transgene expression. This is probably due to gene silencing by a position effect as the R6/0 transgene has clearly integrated into a region of unusual genomic structure (32).

Other mouse models of triplet repeat instability

Triplet repeat instability has also been reported in two series of lines transgenic for the myotonic dystrophy (DM) mutation (CTG on the sense strand) (38,39). The integration fragments used in these lines were a genomic fragment (*Dmt*162) from the myotonic dystrophy (DM) locus containing a small portion of the coding region and the 3′UTR with $(CTG)_{162}$ (39) and a cosmid (DM55-5) containing the myotonic dystrophy protein kinase gene (DMPK) with $(CTG)_{55}$ and the flanking DMR-N9 and DMAHP genes (38). Intergenerational instability was observed in both of these cases. The DM55-5 transgenes showed intergenerational instability in 6.8% of transmissions, the changes generally being expansions of one repeat unit. A higher frequency of unstable transmissions was observed in the *Dmt* lines (as in the R6 lines), most likely as a consequence of the larger size of the repeat tracts.

Table 1. Summary of CAG/polyglutamine transgenic mouse lines in which a progressive neurological phenotype has been observed

Disease	Construct	Promoter	$(CAG)_n$	Expression		Phenotype	Freq. of lines
				RNA	Protein		showing phenotype
HD	exon 1 (genomic)	HD	18	+	+	none	0/2
HD	exon 1 (genomic)	HD	142	–	–	none	0/1
HD	exon 1 (genomic)	HD	115–156	+	+	+	3/3
SCA1	cDNA (full length)	pcp2[a]	30[b]	+	+	none	0/7
SCA1	cDNA (full length)	pcp2[a]	82	+	+	+	5/6
MJD	cDNA (full length)	L7[a]	79	NR	NR	none	0/4
MJD	cDNA (C-terminus)	L7[a]	35	NR	NR	none	0/10
MJD	cDNA (C-terminus)	L7[a]	79	NR	NR	+	3/3
MJD	polyglutamine tract	L7[a]	79	NR	NR	+	2/6

NR, not reported.
[a]Purkinje cell-specific promoter.
[b]Interrupted repeat: $(CAG)_{12}CATCAGCAT(CAG)_{15}$.

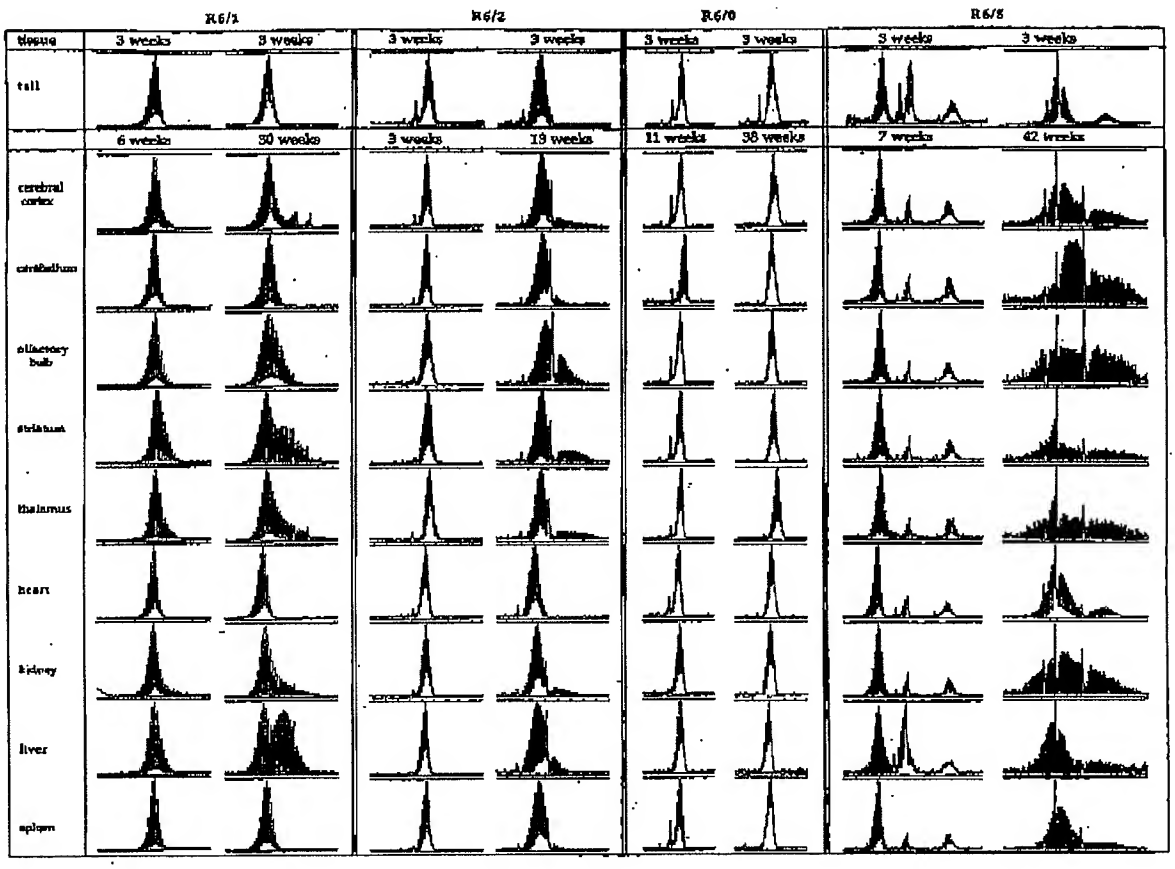


Figure 1. Illustration of the CAG repeat somatic instability seen in the R6 lines. The repeats were amplified by PCR using a fluorescent primer and sized on an ABI sequencer using the Genescan and Genotyper software packages (37). In each case the genescan trace arising from a range of tissues at the age at which the mouse was culled is compared with the trace obtained from tail DNA taken at 3 weeks (top row). The size of the major peaks in the tail traces are: R6/1, 115; R6/2, 145; R6/0, 142; R6/5, range from 123 to 156. The R6/5 line contains four copies of the CAG repeat and the difference in the tail trace between the two R6/5 mice has arisen from germ line instability. It is clear that even after 38 weeks there is no evidence for somatic instability in line R6/0.

The *Dmt* lines [$(CTG)_{143-162}$], like the R6 lines [$(CAG)_{115-155}$], showed a tendency to repeat expansion on male, and contraction on female, transmission. It would appear that the instability seen in the *Dmt* lines parallels that seen in the R6 lines and represents more closely instability seen in some of the CAG/polyGln neurodegenerative disorders rather than that seen in DM. Myotonic dystrophy is caused by a CTG expansion which expands to between $(CTG)_{200}$ and $(CTG)_{4000}$ in the adult and congenital forms of the disease with a maternal bias to the anticipation (40). Somatic instability was described in one of the DM55-5 transgenes which had additional repeat bands in brain, liver, kidney and eye. A similar pattern of instability was also seen in one of the progeny of this mouse, with most instability apparent in sperm (38).

Comparison of the R6, *Dmt* and DM55-5 lines with the transgenic lines that do not exhibit CAG/CTG repeat instability does not lead to an understanding of the molecular basis of instability. The absence of instability observed in the first four reports could not simply be due to a size threshold effect. It is not clear whether the size threshold in the mouse is larger than that seen in humans; however, it must be below 55 repeats as a moderate amount of instability was seen in the DM55-5 transgenes. Similarly, the absence of instability in the first four lines cannot be due to differences in *trans*-acting factors which are likely to be invariant. If *cis*-acting sequences are important, the analysis of four series of lines (30,34–36) which do not show instability and three series (37–39) which fairly consistently do would suggest that these sequences are likely to be present on the

transgenes themselves rather than at the integration sites. The absence of instability in the R6/0 line as compared with the other R6 lines could have mechanistic implications. The R6/0 line only differs from the other three at the site of integration which is probably acting to silence the expression of the transgene. This argues against a model in which contractions and expansions of the repeat occur through a mechanism linked to DNA replication. It raises the possibility that the instability is linked to expression which may be a consequence of the open configuration of chromatin leading to DNA damage rather than being directly linked to transcription.

CONCLUSION

It is clear that a polyGln expansion can give rise to a progressive neurological phenotype in the mouse. The analysis of existing and further transgenic models of CAG/polyGln repeat disease will be informative with respect to uncovering the molecular basis of these disorders. Comparison of transgenes arising from full length and truncated constructs may resolve the speculation that the toxic agent is a truncated version of the proteins in question. The models will be useful in allowing the study of the early disease stages for which patient material is rarely available. Comparison of future models in which the transgenes are under the control of endogenous or ubiquitous promoters may shed light on the factors which determine the differing patterns of neurodegeneration.

REFERENCES

1. HDCRG (1993) A novel gene containing a trinucleotide repeat that is unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983.
2. La Spada, A.R. *et al.* (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79.
3. Koide, R. *et al.* (1994) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.* **6**, 9–13.
4. Nagafuchi, S. *et al.* (1994) Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6**, 14–18.
5. Orr, H.T. *et al.* (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4**, 221–226.
6. Imbert, G. *et al.* (1996) Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nature Genet.* **14**, 285–291.
7. Sampei, K. *et al.* (1996) Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nature Genet.* **14**, 277–284.
8. Pulst, S.-M. *et al.* (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nature Genet.* **14**, 269–276.
9. Kawaguchi, Y. *et al.* (1994) CAG expansions in a novel gene for Machado–Joseph disease at chromosome 14q32.1. *Nature Genet.* **8**, 221–228.
10. Zhuchenko, O. *et al.* (1997) Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage dependent calcium channel. *Nature Genet.* **15**, 62–69.
11. Ross, C.A. (1995) When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. *Neuron* **15**, 493–496.
12. Perutz, M.F. (1996) Glutamine repeats and inherited neurodegenerative diseases: molecular aspects. *Curr. Opin. Struct. Biol* **6**, 848–858.
13. Trottier, Y. *et al.* (1995) Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. *Nature* **378**, 403–406.
14. Li, X.-J. *et al.* (1995) A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378**, 398–402.
15. Wanker, E.E. *et al.* (1997) HIP-1: A huntingtin interacting protein isolated by the yeast two-hybrid system. *Hum. Mol. Genet.* **6**, 487–495.
16. Kalchman, M.A. *et al.* (1996) Huntingtin is ubiquinated and interacts with a specific ubiquitin-conjugating enzyme. *J. Biol. Chem.* **271**, 19385–19394.
17. Burke, J.R. *et al.* (1996) Huntingtin and DRPLA proteins selectively interact with the enzyme GAPDH. *Nature Med.* **2**, 347–350.
18. Goldberg, Y.P. *et al.* (1996) Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nature Genet.* **13**, 442–449.
19. Duyao, M. *et al.* (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genet.* **4**, 387–392.
20. Chung, M. *et al.* (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type 1. *Nature Genet.* **5**, 254–258.
21. Telenius, H. *et al.* (1994) Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature Genet.* **6**, 409–413.
22. Aronin, N. *et al.* (1995) CAG expansion affects the expression of mutant huntingtin in the Huntington's disease brain. *Neuron* **15**, 1193–1201.
23. Chong, S.S. *et al.* (1995) Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **10**, 344–350.
24. Ueno, S. *et al.* (1995) Somatic mosaicism of CAG repeat in dentatorubral-pallidoluysian atrophy (DRPLA). *Hum. Mol. Genet.* **4**, 663–666.
25. Takano, H. *et al.* (1996) Somatic mosaicism of expanded CAG repeats in brains of patients with dentatorubral-pallidoluysian atrophy: cellular population-dependent dynamics of mitotic instability. *Am. J. Hum. Genet.* **58**, 1212–1222.
26. Tanaka, F. *et al.* (1996) Differential pattern of tissue-specific somatic mosaicism of expanded CAG trinucleotide repeat in dentatorubral-pallidoluysian atrophy, Machado–Joseph disease, and X-linked recessive spinal and bulbar muscular atrophy. *J. Neurol. Sci.* **135**, 43–50.
27. Nasir, J. *et al.* (1995) Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell* **81**, 811–823.
28. Duyao, M.P. *et al.* (1995) Inactivation of the mouse Huntington's disease gene homolog *Hdh*. *Science* **269**, 407–410.
29. Zeitlin, S. *et al.* (1995) Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nature Genet.* **11**, 155–163.
30. Goldberg, Y.P. *et al.* (1996) Absence of the disease phenotype and intergenerational stability of the CAG repeat in transgenic mice expressing the human Huntington's disease transcript. *Hum. Mol. Genet.* **5**, 177–185.
31. Hodgson, J.G. *et al.* (1996) Human huntingtin derived from YAC transgenes compensates for loss of murine huntingtin by rescue of the embryonic lethal phenotype. *Hum. Mol. Genet.* **5**, 1875–1885.
32. Mangiarini, L. *et al.* (1996) Exon 1 of the Huntington's disease gene containing a highly expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* **87**, 493–506.
33. Davies, S.W. *et al.* (1997) Formation of neuronal intranuclear inclusions (NII) underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell*, in press.
34. Burright, E.N. *et al.* (1995) SCA1 transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. *Cell* **82**, 937–948.
35. Ikeda, H. *et al.* (1996) Expanded polyglutamine in the Machado–Joseph disease protein induces cell death *in vitro* and *in vivo*. *Nature Genet.* **13**, 196–202.
36. Bingham, P.M. *et al.* (1995) Stability of an expanded trinucleotide repeat in the androgen receptor gene in transgenic mice. *Nature Genet.* **9**, 191–196.
37. Mangiarini, L. *et al.* (1997) Instability of highly expanded CAG repeats in transgenic mice is related to expression of the transgene. *Nature Genet.* **15**, 197–200.
38. Gourdon, G. *et al.* (1997) Moderate intergenerational and somatic instability of a 55-CTG repeat in transgenic mice. *Nature Genet.* **15**, 190–192.
39. Monckton, D.G. *et al.* (1997) Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nature Genet.* **15**, 193–196.
40. Wieringa, B. (1994) Commentary: Myotonic dystrophy reviewed: back to the future? *Hum. Mol. Genet.* **3**, 1–7.

EXHIBIT 9

Cell, Vol. 87, 493–506, November 1, 1996, Copyright ©1996 by Cell Press

Exon 1 of the *HD* Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice

Laura Mangiarini,[1] Kirupa Sathasivam,[1] Mary Seller,[1] Barbara Cozens,[7] Alex Harper,[2] Colin Hetherington,[3] Martin Lawton,[4] Yvon Trottier,[5] Hans Lehrach,[6] Stephen W. Davies,[7] and Gillian P. Bates[1]
[1]Division of Medical and Molecular Genetics
UMDS
Guy's Hospital
London SE1 9RT
United Kingdom
[2]UMDS Transgenic Unit
The Rayne Institute
St. Thomas's Hospital
London SE1 7EH
United Kingdom
[3]Biomedical Services
John Radcliffe Hospital
University of Oxford
Oxford OX3 9DU
United Kingdom
[4]Biological Services Division
UMDS
Guy's Hospital
London SE1 9RT
United Kingdom
[5]Institut de Genetique et Biologie Moleculaire et Cellulaire
CNRS/INSERM/ULP
Illkirch
CU Strasbourg 67404
France
[6]Max Planck Institut fur Molekulare Genetik
Dahlem, Berlin D14195
Germany
[7]Department of Anatomy and Developmental Biology
University College
London WC1E 6BT
United Kingdom

Summary

Huntington's disease (HD) is one of an increasing number of neurodegenerative disorders caused by a CAG/polyglutamine repeat expansion. Mice have been generated that are transgenic for the 5′ end of the human *HD* gene carrying $(CAG)_{115}$–$(CAG)_{150}$ repeat expansions. In three lines, the transgene is ubiquitously expressed at both mRNA and protein level. Transgenic mice exhibit a progressive neurological phenotype that exhibits many of the features of HD, including choreiform-like movements, involuntary stereotypic movements, tremor, and epileptic seizures, as well as nonmovement disorder components. This transgenic model will greatly assist in an eventual understanding of the molecular pathology of HD and may open the way to the testing of intervention strategies.

Introduction

Huntington's disease (HD) is an autosomal dominant progressive neurodegenerative disorder (Harper, 1991). The onset of symptoms is generally in midlife although it can range from early childhood to >70 years. Anticipation is observed, predominantly when the disease is inherited through the male line, with the result that 70% of juvenile cases inherit the disease from their father. The symptoms have an emotional, motor, and cognitive component. A detailed description of all aspects of HD can be found in Harper (1991). Chorea is a characteristic feature of the motor disorder and is defined as excessive spontaneous movement, irregularly timed, randomly distributed, and abrupt. It can vary from being barely perceptible to extremely severe. It involves all parts of the body, can have repetitive and stereotypic elements, and may have a pseudopurposive appearance (Harper, 1991). Other frequently observed motor abnormalities include dystonia (sustained muscle contraction), rigidity, bradykinesia (abnormally slow movements), oculomotor dysfunction, and tremor. Cerebellar dysfunction, upper motorneuron abnormalities, epilepsy, and myoclonus (brief shock-like muscle jerks) are rare except in the juvenile form of the disease, which commonly presents with a "Parkinsonlike rigidity." Voluntary movement disorders include fine motor incoordination, dysarthria (impairment of articulation), and dysphagia (difficulty in swallowing). The emotional disorder is commonly depression and irritability, and the cognitive component comprises a subcortical dementia. The biochemical basis of this disease is not understood, and there is no effective therapy.

The HD mutation results in the expansion of a polyglutamine (polygln) tract in a large 350 kDa protein of unknown function (Huntington's Disease Collaborative Research Group, 1993). The normal and expanded *HD* allele sizes have been defined as CAG_{6-37} and CAG_{35-121} repeats, respectively. An inverse correlation between age of onset and repeat length is most pronounced for juvenile HD for which the longest repeats have been observed (Huntington's Disease Collaborative Research Group, 1993; Telenius et al., 1993). Despite the selective cell death, the *HD* transcript is ubiquitously expressed (Strong et al., 1993). The polyglutamines are successfully translated and the huntingtin protein (htt) products arising from expanded alleles have been identified in protein extracts from HD patients (Jou and Myers, 1995; Trottier et al., 1995a).

CAG/gln expansion has been found to be the causative mutation in five neurodegenerative diseases for which the gene has been cloned. In addition to HD, these include spinal and bulbar muscular atrophy (SBMA) (La Spada et al., 1991), spinocerebellar ataxia type 1 (SCA1) (Orr et al., 1993), dentatorubral-pallidoluysian atrophy (DRPLA) (Koide et al., 1994), and Machado Joseph disease (MJD or SCA3) (Kawaguchi et al., 1994). Many aspects of the genetics and molecular biology are common to these diseases. They are autosomal dominant (with the exception of X-linked SBMA) and show varying degrees of anticipation on paternal transmission. The size of the normal and expanded CAG repeat ranges are comparable, and available data indicate that age of onset correlations and patterns of repeat stability are

reproduced. A similar ubiquitous expression pattern is also characteristic, and the presence of the expanded forms of ataxin-1 (SCA1 protein) and atrophin-1 (DRPLA protein) in lysates from patient tissues have been observed (Servadio et al., 1995; Yazawa et al., 1995).

Despite the otherwise apparent universality of this mutation, the patterns of cell death differ between these diseases. In HD, the most striking atrophy occurs in the caudate nucleus, which is often reduced to a rim of tissue. The putamen and globus pallidus also undergo atrophy, and there are subtle changes in the cerebral cortex (Vonsattel et al., 1985). SBMA is a form of motor neuron disease with both spinal and bulbar motor neuron involvement (Kennedy et al., 1968). The SCA1 and SCA3 spinocerebellar ataxias are clearly distinguished by major neuropathological features: Purkinje cell, pontine nuclei, and inferior olivary nuclei degeneration in SCA1 (Zoghbi et al., 1993) and pontine nuclei and the molecular layer of the cerebellum in SCA3 (Durr et al., 1996). In DRPLA, neuropathology includes the cerebellar dentate nucleus, globus pallidus, red and subthalamic nuclei, Purkinje cells, brain stem tegmentum, and the lateral corticospinal tract (Takahashi et al., 1988). The proteins containing the polygln repeats are otherwise unrelated. In SBMA, the repeat lies within the androgen receptor (La Spada et al., 1991), while the others are in novel genes of unknown function. Subcellular localization suggests differing roles for these proteins (DiFiglia et al., 1995; Servadio et al., 1995; Trottier et al., 1995a; Yazawa et al., 1995).

It is essential that transgenic models of these diseases are developed. There have been two previous reports of a neurological phenotype observed in mice transgenic for a protein carrying a polygln repeat expansion. The first used a *SCA1* cDNA construct with $(CAG)_{82}$ under the control of a Purkinje cell specific promoter (Burright et al., 1995). Three heterozygous lines overexpressing the *SCA1* transcript by 10- to 100-fold and two homozygous lines showed a progressive ataxic phenotype between 12 and 26 weeks of age. The mice became clearly ataxic when walking and routinely fell when attempting to stand on their hind legs. Pathologic examination showed significant loss of the Purkinje cell population with Bergmann glial proliferation and shrinkage and gliosis of the molecular layer. More recently, transgenic mice have been reported with a $(CAG)_{79}$ version of the *SCA3* gene and also the $(CAG)_{79}$ polygln tract in isolation, both under the control of the Purkinje cell specific promoter (Ikeda et al., 1996). Affected mice transgenic for the isolated polygln tract were severly ataxic, they exhibit a wide-based hind limb stance, frequently fall when moving, and are unable to rear. Overt Purkinje cell death was observed with secondary effects to the molecular and granular cell layers. No phenotype was observed in the mice transgenic for the entire mutated *SCA3* gene. The authors suggested that the polygln tracts are more toxic in isolation than in the context of a protein, although in the absence of any information concerning transgene copy number, genomic structure of the integration sites, or expression levels, this interpretation should be treated with caution. These reports have shown that Purkinje cell specific overexpression of an expanded polygln tract, both in the context of the *SCA1* gene or in isolation, is toxic to Purkinje cells and causes a corresponding ataxic phenotype.

In our intitial attempt to generate a murine model of HD, we have focused on the construction of a mutant yeast artificial chromosome (YAC) for introduction by pronuclear injection. Progress was severely hampered by both instability of YAC intermediates and the severe instability of highly expanded CAG repeats in yeast. Consequently, to address the question of CAG repeat stability in the mouse, transgenic lines were established with a 1.9 kb human genomic fragment containing promoter sequences and exon 1 carrying expansions of approximately $(CAG)_{130}$. Unexpectedly, this fragment has been sufficient to generate a progressive neurological phenotype that displays many of the characteristics of HD. This is the first time that a model of one of these diseases has been generated by a transgene driven from an endogenous promotor. The availability of a mouse model of the disease is extremely informative with regard to the size of the polygln expansion and level of expression required to produce a phenotype with a given age of onset in the mouse. This work suggests that the polygln-containing domain of the htt protein may be sufficient to generate a mouse model of HD.

Results and Discussion

Fragment Used for Transgenesis

The microinjection fragment was a 1.9 kb SacI–EcoRI fragment from the 5′ end of the human *HD* gene isolated from a phage genomic clone derived from an HD patient (Figure 1a). It is composed of ~1 kb of 5′ UTR sequences, exon 1 carrying expanded CAG repeats of ~130 units and the first 262 bp of intron 1. As the CAG repeats are unstable when propagated in E. coli, the DNA preparation used for microinjection contained a heterogeneous set of repeats of varying size but of the order of 130 units. In the event that an unspliced mRNA should be transcribed from this fragment, an "in-frame" stop codon immediately at the beginning of intron 1 would result in a truncated protein corresponding to the first 90 amino acids of the published htt protein (repeat size of $(CAG)_{21}$).

Genomic Organization of the Integration Events

Transgenic mice were generated by microinjection of single cell CBAxC57BL/6 embryos. Of 29 newborn mice, seven died neonatally, and of the remaining 22 pups, one male was transgenic. This founder (R6) was initially backcrossed to both C57BL/6 and to CBAxC57BL/6 females. However, a subsequent need to optimize litter size has resulted in the maintenance of the transgene on the CBAxC57BL/6 hybrid background. F1 mice were genotyped both by Southern analysis and by PCR to determine the CAG repeat size. Figure 1b shows a Southern blot of BamHI digested DNA from a number of F1 progeny. It was possible to deduce that the microinjection fragment had integrated into five different regions of the founders' genome. The predicted genomic organization of the integration events is illustrated in Figure 1c. In lines R6/1 and R6/0, the fragment has integrated as an intact single copy, and in line R6/T as

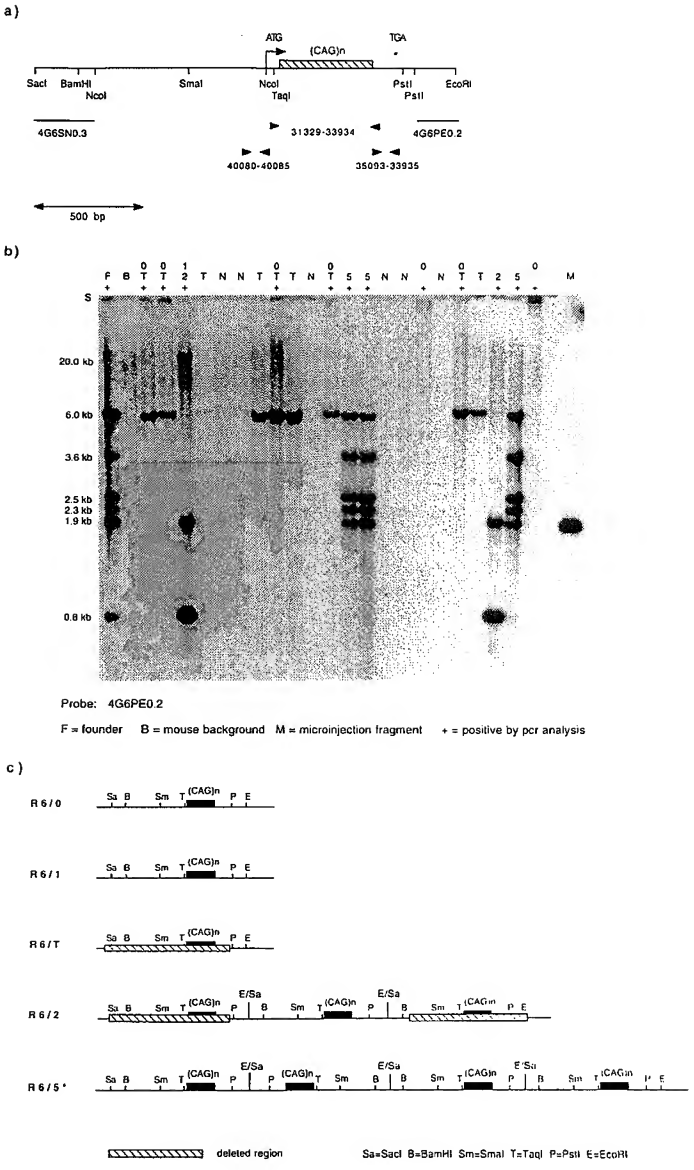


Figure 1. Microinjection Fragment and Identification of the Integration Events

(a) Restriction map of the human genomic fragment used for microinjection. An arrow denotes the transcription start site and asterisk indicates the position of an in-frame stop codon at the beginning of intron 1. 4G6SN0.3 and 4G6PE0.2 are fragments used as hybridization probes, and solid triangles indicate the location of PCR assays used for genotyping and RNA analysis.

(b) Southern blot of genomic DNA from the R6 founder and a number of F1 progeny. DNA was digested with BamHI and probed with 4G6PE0.2. The genotypes are indicated above the lanes (1, 2, 0, T, or 5). A plus sign indicates that the mouse also scored as transgenic when typed with the CAG repeat PCR assay. BamHI fragment sizes are as follows: R6/1, 20.0 kb; R6/2, 1.9 and 0.8 kb; R6/5, 6.0, 3.6, 2.5, 2.3, and 1.9 kb; R6/0, band migrates close to slot (S); R6/T, 6.0 kb. The R6/T genotype is negative with the CAG repeat PCR assay.

(c) Genomic organisation of the integration sites of the transgenes. R6/0, R6/1, and R6/T are single copy integrants although R6/T is highly deleted. R6/2 probably originated as a three copy integrant, the flanking fragments having undergone deletions. (asterisk) It has not been possible to completely resolve the structure of the R6/5 integration event. Three of the five BamHI fragments can be accounted for by the structure as drawn.

a highly truncated fragment. In line R6/0, the fragment has most probably inserted adjacent to a repetitive genomic structure. When the probe 4G6PE0.2 is hybridized to Southern blots of transgene genomic DNA digested with BamHI, SmaI, PstI, or NcoI, in each case a band is detected that has barely migrated into the gel. If the same blots are probed with 4G6SN0.3, the 5′UTR probe, bands of a more expected size range are seen. Line R6/2 most probably originated as a three copy integration event, the flanking fragments having been subject to deletions, with the result that this transgene functions essentially as a single copy integrant. Finally line R6/5 is represented by five bands on a BamHI Southern blot. It is clear that four fragments have integrated as illustrated in Figure 1c. This includes both a tail-to-tail and head-to-head arrangement. However, other hybridization bands could not be explained by a straightforward configuration, as in those illustrated, or by simple deletions. It seems likely, therefore, that a complicated rearrangement must have occurred for which it has not been possible to completely unravel the genomic structure.

Size of the CAG Expansion in Each of the Transgenic Lines

Four of the transgenic lines: R6/0, R6/1, R6/2, and R6/5 carry expanded CAG repeats. The size of the expansion was determined by PCR amplification of the repeat using a fluorescently labeled primer and subsequent size determination using an ABI sequencer (Figure 2). The peak sizes are as follows: R6/1, 116 repeat units; R6/0, 142 repeat units; R6/2, 144 repeat units. Line R6/5 is more complicated with peaks at 128, 132, 135, 137,

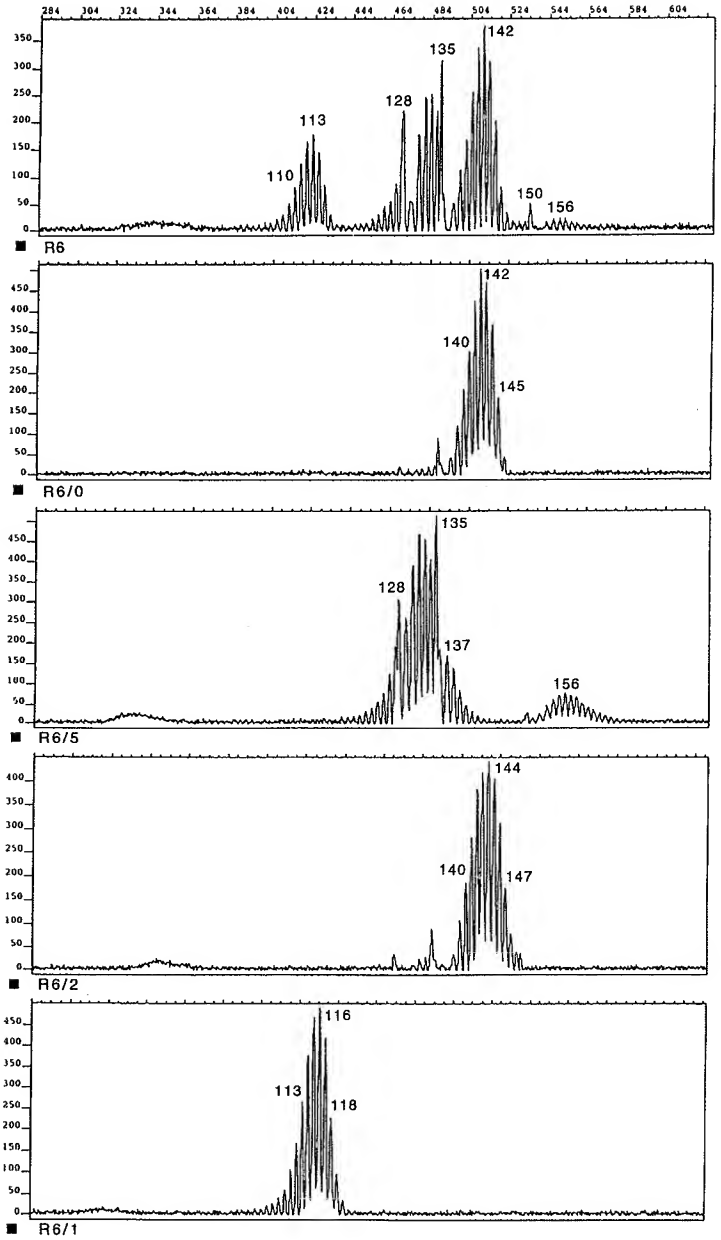


Figure 2. Measurement of the Size of the CAG Expansion in the R6 Transgenic Lines

The CAG repeats were amplified with a FAM-labeled primer as described. The top panel shows the trace specific to the founder (R6) and the four lower panels, the traces obtained in F1 mice with the R6/0, R6/5, R6/2, and R6/1 genotypes.

and 156 repeat units. These repeat sizes are considerably larger than those that have generally been reported to cause the juvenile form of HD in humans. Both gametic and somatic repeat instability have been observed (manuscript submitted).

Segregation of the Integration Events

The specific genotype frequencies found in 321 F1 mice derived from the R6 founder are summarized in Table 1. The integration events appear to segregate independently but are only seen in certain combinations. The founder is therefore a germ line chimera with one set of germ cells containing the R6/0 and R6/T transgenes and the other containing the R6/1, R6/2 and R6/5 transgenes.

Phenotype Observed in the R6/2 Transgenic Line

The age of onset in line R6/2 has been observed as early as four weeks (one mouse) but most frequently occurs between nine and eleven weeks. Age at death has generally been between 10 and 13 weeks although the mouse with the age of onset at four weeks died at six and a half weeks. The mice display a progressive neurological phenotype. As far as can be ascertained, the mice remain alert, exploratory and inquisitive, and responsive

Table 1. Frequency of Genotypes Arising in 321 F1 Progeny

Genotype	N
R6/0	56
R6/T	61
R6/0 + R6/T	62
Negative[a]	(56/71)
R6/1	8
R6/2	16
R6/5	15
R6/1 + R6/2	9
R6/1 + R6/5	8
R6/2 + R6/5	11
R6/1 + R6/2 + R6/5	4
Negative[a]	(15/71)

[a] The total number of nontransgenic mice are divided between the two genotype clusters in a proportion consistent with the genotype frequencies.

to sensory stimuli. The phenotype is complex. There are a number of components to the motor disorder including a resting tremor, movements described as resembling chorea, stereotypic involuntary movements, and in some cases a mild ataxia manifesting as dysmetria. One of the first symptoms is a dyskinesia of the limbs when held by the tail. This progresses to an alternating clasping together and releasing of the feet until the mice clasp their feet together immediately after they are picked up, (Figure 3a), and can no longer release this posture. The mice develop a constant tremor that becomes progressively worse. The tremor tends to be less noticeable when they are quiet or asleep, but worsens under stress (for example, the removal of the cage lid) or if they reach for food or to climb out of the cage. As the disorder progresses, stereotypic involuntary movements are apparent, which include repetitive stroking of the nose and face, and a hind limb kicking/scratching motion. Sudden movements that involve the whole body and may resemble chorea are observed. These are rapid, abrupt, irregular, and manifest as a shaking/shudder of the trunk. The mice do not develop a wide-based gait, can stand on their hind limbs and climb out of the cage without falling. They only consistently lose balance when sitting on their hind limbs, turning, and reaching round to groom their backs, which results in a somersault. The mice exhibit severe handling-induced epileptic seizures that can last for several minutes.

At weaning, the R6/2 transgenes are indistinguishable from their normal litter mates. Coincident with the onset of motor symptoms, their weight plateaus and then progressively decreases. In the end stages, mice have been observed to weigh as little as 60%–70% of their normal sibs. As the disease becomes more severe, they are very frequently observed to be eating but do not gain weight. It appears that the mice are eating rather than just breaking off food. Their food comprises an expanded chow, which does not crumble easily, and excess food crumbs are not observed in the bedding. On autopsy, the mice are often emaciated with an overall loss of muscle bulk although food is observed in the stomach and fecal pellets in the gut. Histological analysis of muscle samples showed no evidence of a myopathy.

Characteristic vocalizations have been observed. These include a sound similar to that made by a new born litter, which resembles teeth chattering from cold, but is likely to have a respiratory basis (since it occurs before the young mice have teeth). A second sound, a type of chirping noise, is more reminiscent of a bird than of a mouse. The mice are more likely to make these sounds when they are under stress (for example, away from the home cage).

The mice appear to urinate more frequently. The bedding at the front of the cage becomes excessively wet as compared to that in cages housing normal mice. They are unlikely to be suffering from spastic bladders as the wetting of the bedding is not uniform. Urine tests in 18 transgenic mice (11 male and 7 female) showed no abnormality in glucose or protein levels. Similarly, blood tests in two mice showed glucose and protein levels to be within the normal range.

R6/2 females are sterile and, of ten R6/2 males that have been placed with females from a time just prior to expected sexual maturity, five have mated. Of these, one mouse produced one litter, two mice produced two litters and two produced four litters. On autopsy, the reproductive organs consistently appear vestigial or atrophied. Females often have miniscule ovaries and a hair-like uterus. Males have small testes, seminal ducts, and coagulation glands. On histology, one male that had failed to mate was found to have testicular atrophy with an absence of spermatazoa, an atropy of the epididymus with aspermia, and no secretion present in the coagulation gland.

The mice die suddenly and the cause of death is generally unknown although one mouse was observed to die during an epileptic seizure.

Dosage Effect on Age of Onset and Phenotype Severity in Complex Genotypes

Lines R6/1, R6/2, and R6/5 have been established from the founder. In the F1 generation, mice with all possible combinations of these transgenes were identified. Each aspect of the phenotype, as described for line R6/2, has been observed for the genotypes listed in Table 2. In the end stages of the disease, the transgenes are always considerably smaller than their normal littermates. The age of onset varies from <3 weeks (R6/1+R6/2+R6/5 genotype) to ~4 or 5 months (R6/1 line).

The (R6/1+R6/2+R6/5) genotype is the most severe. Only four such mice were recovered in the F1 generation. The overall genotype frequency (Table 1) would have predicted more than this, and it is possible that some mice with this genotype died neonatally or in utero. All aspects of the phenotype are more severe and have a more rapid progression. The (R6/1+R6/2+R6/5) mice are considerably smaller than their litter mates at weaning. For example, one weighed 5.2 g at 23 days of age as compared to a mean of 9.2 g for her female sibs. She reached a maximum weight of 7.5 g but was only 6.0 g at death at 51 days as compared to a mean of 16.3 g for her sibs. In contrast, line R6/1 has the latest age of onset and the slowest progression. The mice begin to exhibit the feet-clasping posture when suspended by the tail at ~4–5 months. At between 6 and 7 months,

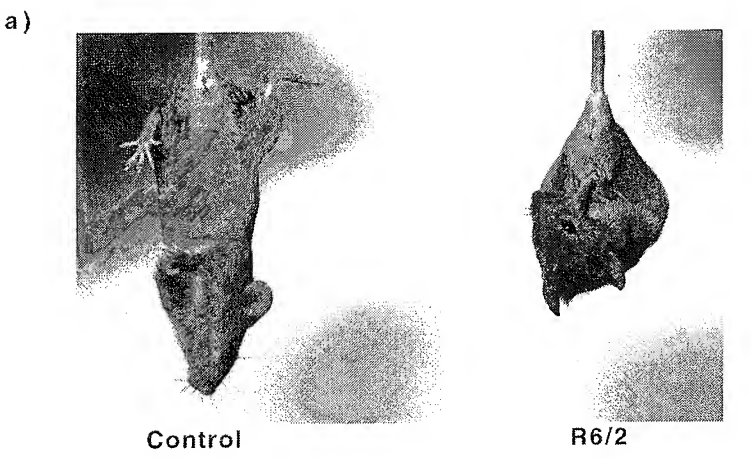


Figure 3. Comparison of R6 Transgenic Mice and Littermate Controls

(a) An R6/2 transgenic mouse demonstrating the feet-clasping posture adopted when suspended by the tail. The normal mouse holds its hind limbs outward in order to steady itself.
(b) The R6/2 mouse (17.7 g) and normal littermate (21.3 g) at 12 weeks of age. The transgenic mouse is thinner.
(c) An R6/1+R6/2 (10.1 g) transgenic mouse and normal littermate (19.6 g) at seven weeks, three days. There is a considerable size difference.

some show a mild tremor and intermittently exhibit all aspects of the involuntary movement disorder as described for the R6/2 line. Epileptic seizures have also been observed. The effect of transgene dosage on the size of the mice is illustrated in Figure 3.

On autopsy, atrophy or gross atrophy of the primary and secondary reproductive organs is routinely observed. Otherwise, hepatic changes in the form of polyploid hepatic nuclei and a loss of cytoplasmic mass with no obvious cell death was the only consistent observation resulting from a routine histopathological examination (two R6/2 and six R6/1+R6/5 mice in the end stages of the disease and displaying all aspects of the phenotype). Thymic atrophy is sometimes present, more frequently in the more severely affected lines, but this does not correlate with the presence or absence of phenotypic features. In a few mice there is a slight deformation to the cranial vault resulting in a boney ridge over the

Table 2. Comparison of the Onset and Duration of the Phenotype Associated with the R6 Genotypes

	Age of Onset	Age at Last Litter	Age at Death
R6/1 + R6/2 + R6/5	< 3 weeks	N/A	4–7 weeks
R6/1 + R6/2	3–4 weeks	N/A	6–8 weeks
R6/2 + R6/5	6–7 weeks	N/A	8–12 weeks
R6/2	9–11 weeks	6–9 weeks[a] (5 males)	10–13 weeks
R6/1 + R6/5	12–16 weeks	12 weeks[a] (1 male)	24–36 weeks
R6/1	15–21 weeks	14 weeks[b] (1 male)	32–40 weeks[c]

[a] Mice bred continuously.
[b] Mouse failed to breed when cross set up at 19 weeks.
[c] Oldest R6/1 mouse is alive at 40 weeks.

cerebellum. This has been seen more frequently in the lines with the more severe phenotype but has also been observed in line R6/1+R6/5.

A phenotype has not been observed in the heterozygous (R6/5)/+ or (R6/0)/+ lines, the oldest mice now being ~14 months. R6/5 homozygotes are developing symptoms at ~9 months, and the R6/5 transgene clearly contributes to the onset and progression of the disorder when in combination with R6/1 or R6/2 transgenes.

Expression of the Transgene

PCR primers specific to exon 1 of the human *HD* gene were used to examine the expression and tissue distribution of the transgenes. RT–PCR showed the transgene to be expressed in every tissue examined for lines R6/2 (Figure 4a), R6/1, and R6/5, but was not expressed in line R6/0. This ubiquitous pattern of expression for three of the lines suggests that the transgene is most likely expressed from promoter sequences present on the microinjection fragment. The absence of expression in line R6/0 is probably due to a position effect as Southern analysis of this line predicts that the R6/0 transgene has integrated adjacent to a genomic region of unusual structure. Northern analysis revealed transcripts of 2.5 and 2.3 kb in lines R6/1 and R6/2, respectively (Figure 4b) and the suggestion of a larger R6/5 transcript. The 4G6PE0.2 probe is derived from intron 1 of the human gene, and the presence of this sequence in the transcripts indicates that the human exon 1 has not spliced to mouse exonic sequences potentially occurring close to the integration sites.

The level of expression of the transgene with respect to the endogenous mouse *hd* gene was assessed in total RNA from six tissues for each of the lines R6/1, R6/2, R6/5, and R6/0. The PCR primers had identical recognition sequences in exon 1 of both the mouse and human genes and amplified mouse and human products of 121 and 114 bp, respectively. No expression was detected in the R6/0 transgene. While the comparative expression level varies between tissues, the average expression of the R6/2, R6/1, and R6/5 transgenes was 75%, 31%, and 77% of the endogenous level (data not shown). The tissue variability made absolute quantitation difficult, but this analysis nevertheless places the level of expression of the transgene within the range of the murine gene.

The monoclonal antibody, 1C2, binds specifically and in a size-dependent manner to pathogenic polygln expansions (Trottier et al., 1995b). This antibody was used to immunoprobe Western blots of cell lysates derived from a complete set of tissues from lines R6/1, R6/2, and R6/5. A transgene-specific product was detected in lines R6/2 and R6/5 in all tissues tested. Figure 5 shows the Western blots obtained for a subset of tissues from lines R6/2 and R6/5. The predicted size of the R6/2 protein would be ~23 kDa. The migration of the R6/2 and R6/5 products, at a size larger than this with respect to the markers, is consistent with the aberrant migration observed for the expanded polygln containing htt, ataxin-1, and atrophin-1 products when compared to their normal counterparts. A constant band detected in all transgene and control tissues was found to be due to cross-reactivity of the antimouse secondary antibody. Comparison of the intensity of the the constant band between the R6/2 and R6/5 tissues suggests that the transgene protein is present at similar levels in these lines. A protein product has not been detected in line R6/1 despite testing ranges of polyacrylamide concentration and antibody dilution. It would be extremely unlikely that a protein product were not present in this line. One possible explanation is that the length of polygln tract in the R6/1 protein does not present an epitope to the 1C2 antibody. It is not clear from expression analysis why the R6/5 phenotype should be so much milder than that observed in lines R6/2 and R6/1.

Neuropathology

Nine R6/2 transgenic mice, exhibiting a broad spectrum of severe symptoms of 2–3 weeks duration, and nine nontransgenic littermates were used for neuropathological investigation. Brains from the transgenic animals were consistently smaller than controls (controls 490 ± 9.8 mg, transgenes 395 ± 8.0 mg). Serial 40 μm sections in either the coronal (12 mice) or horizontal (6 mice) planes were processed for either Nissl staining (Figures 6 and 7) or the immunocytochemical localization of glial fibrillary acidic protein (GFAP) or the mouse macrophage and microglial marker F4/80. The morphology of the central nervous system (CNS) in the transgenic mice appeared normal with no focal areas of malformation or neurodegeneration; however, sections of the brains of these animals were consistently smaller than those of their litter mates (19% ± 1.6%). This reduction in size appeared to be uniform throughout all CNS structures. Analysis of thionin-stained sections showed no evidence of neuronal cell loss, oligodendrocyte loss, reac-

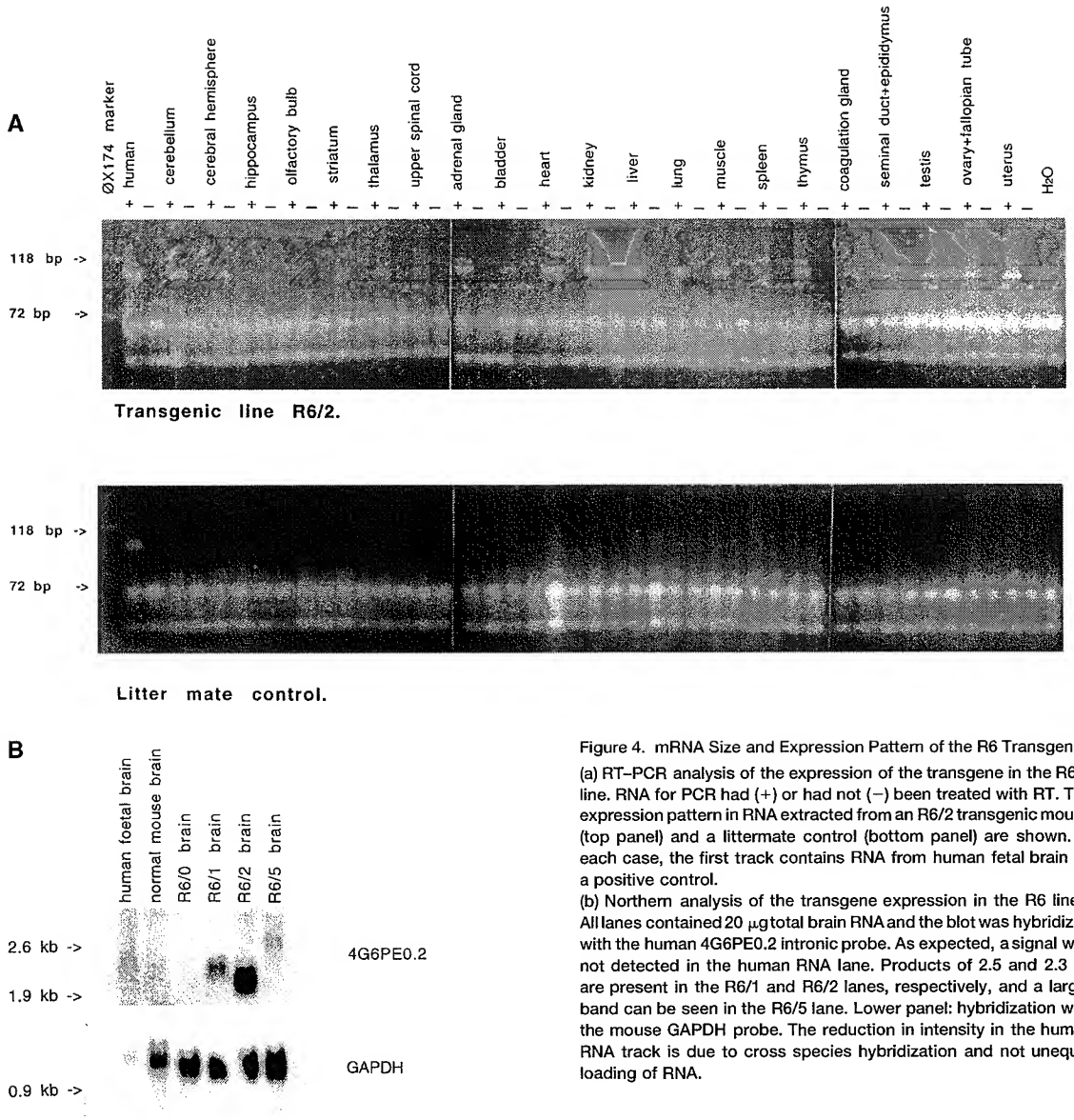


Figure 4. mRNA Size and Expression Pattern of the R6 Transgenes

(a) RT–PCR analysis of the expression of the transgene in the R6/2 line. RNA for PCR had (+) or had not (−) been treated with RT. The expression pattern in RNA extracted from an R6/2 transgenic mouse (top panel) and a littermate control (bottom panel) are shown. In each case, the first track contains RNA from human fetal brain as a positive control.

(b) Northern analysis of the transgene expression in the R6 lines. All lanes contained 20 μg total brain RNA and the blot was hybridized with the human 4G6PE0.2 intronic probe. As expected, a signal was not detected in the human RNA lane. Products of 2.5 and 2.3 kb are present in the R6/1 and R6/2 lanes, respectively, and a larger band can be seen in the R6/5 lane. Lower panel: hybridization with the mouse GAPDH probe. The reduction in intensity in the human RNA track is due to cross species hybridization and not unequal loading of RNA.

tive gliosis, or inflammatory change. These latter two observations were corroborated by the GFAP and F4/80 stained sections, where the normal distribution of astrocytes and ramified microglia cells was observed in the absence of any indication of increased reactivity of astrocyte staining or the presence of rounded microglia or infiltrating macrophages.

Cerebral Cortex and Hippocampus

The cytoarchitectonic structure of the cerebral cortex was maintained in the frontal, temporal, occipital and parietal lobes, although all regions were noticeably thinner when measured between the pia and subcortical white matter. The large pyramidal cells of the motor regions of the frontal cortex were present in normal number and morphological appearance. Similarly the pyramidal cells of hippocampus, subiculum and parahippocampal gyrus, the stellate cells of layer II of the entorhinal cortex, and the granule cells of the dentate gyrus were of normal size and distribution.

Basal Ganglia

A detailed analysis of the striatum, nucleus accumbens, globus pallidus, entopenduncular nucleus, subthalamic nucleus, and substantia nigra demonstrated normal neuronal density and patterns of morphological diversity. The striatum is composed of a normal complement of medium-sized striatal neurons interspersed with fewer large and small neurons, together with satellite glia. The white matter of the corpus callosum and the fascicles of fibers forming the internal capsule contain as many oligodendrocytes as similar sections from control mice. The striatum is again consistently smaller in the transgenic animals.

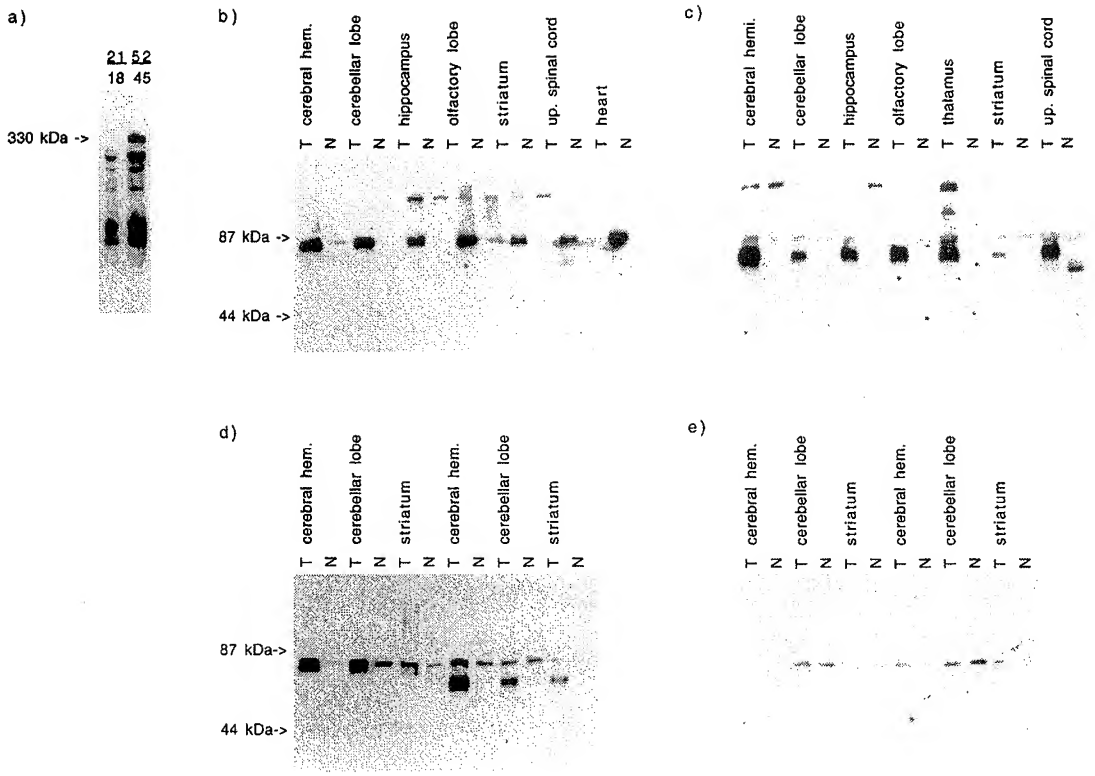


Figure 5. Expression Profile of the Transgene Protein Products

Identification of the transgene protein product in the R6/2 and R6/5 lines using a monoclonal antibody (1C2) that specifically detects polygln expansions.

(a) Identification of htt in lysates prepared from lymphoblastoid cell lines from a normal individual and an HD homozygote and fractionated on a 6% SDS–PAGE gel. The size of the respective CAG expansions are indicated above the tracts. The position at which the fibrinogen marker (330 kDa) migrates is indicated.

(b) Lysates from an R6/2 transgene (T) and littermate control (N) were fractionated on a 10% SDS-PAGE gel.

(c) Lysates from an R6/5 transgene (T) and a litermate control were fractionated on a 10% gel.

(d) Lysates from the R6/2 and R6/5 lines fractionated on a 10% SDS–PAGE gel.

(e) The filter in (d) stripped and reprobed with the secondary antimouse antibody, which detects the constant band seen in (b)–(d).

Cerebellum and Spinal Cord

The granule cells, Purkinje cells, and the neurons of the molecular layer of the cerebellum show no differences from the control mice. Similarly, the large motor neurons of the anterior horn of the cervical and lumbar enlargements of the spinal cord and the dorsal horns are again of normal appearance.

Examination of all other areas of the CNS revealed no gross or microscopic abnormalities.

Discussion

Transgenic mice that develop a progressive neurological phenotype have been generated by the introduction of a genomic fragment containing exon 1 of the human *HD* gene. Four lines have been established, with CAG repeat expansions ranging from ~115 to 150 repeat units. In the three lines that exhibit a phenotype, R6/1, R6/2, and R6/5, the transgene has a ubiquitous mRNA and protein expression pattern. The transgene mRNA is most likely transcribed from human promoter elements and extends into the flanking mouse sequences. The presence of human intron 1 sequences in the mRNA rules out the possibility that the human exon splices to mouse exonic sequences and therefore predicts that the corresponding transgene protein products contain 69 amino acids in addition to the number of polygln residues encoded by the repeat expansion.

The polygln expansions in the R6 transgenic mice are of a size considerably greater than is generally associated with the juvenile form of HD. Even so, it is not possible to predict the phenotypic expression of such a mutation in the mouse. In HD, the major focus of neuropathological change is in the striatum (part of the basal ganglia) and the cerebral cortex. The motor disorder observed in the R6 lines is strongly suggestive of a basal ganglia lesion. The mice exhibit involuntary jerky shudders that have been described as resembling chorea and likened to the choreic movements observed in the neurological disease arising from canine distemper (Lauder et al., 1954). As far as we can ascertain, chorea has not previously been described in mice (Lyon and Searle, 1990). The neuropathological correlate of chorea is accepted as a basal ganglia lesion. The pronounced

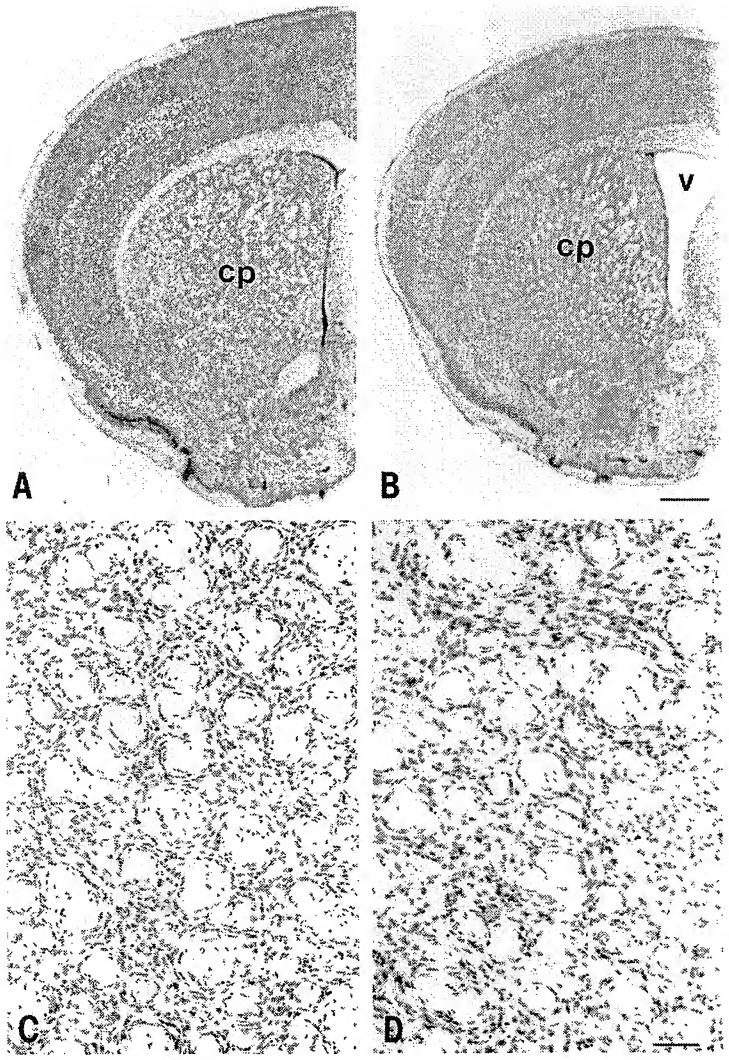


Figure 6. Nissl Sections Through the Mouse Forebrain

Frontal section through the caudate/putamen (cp) at the level of the lateral ventricle (v) of a normal littermate control (A) and R6/2 transgenic mouse (B). The caudate putamen is shown in higher power in C (control) and D (R6/2 transgene). Scale bars, 500 μm in (A) and (B), 80 μm in (C) and (D).

progressive resting tremor that occurs in all limbs, trunk, and head of affected mice also points to a basal ganglia abnormality. The observation of epileptic seizures is compatible with juvenile HD; however, while seizures have a cerebral focus, they could result from many imbalances that are both intracranial or extracranial.

The R6 mice also suffer from a progressive decrease in body weight and an overall loss of muscle bulk. Similarly, loss of body weight and a generalized lack of muscle bulk is a progressive and characteristic symptom of HD, despite increased calorific intake (Sanberg et al., 1981). The weight loss appears to be independent of the hyperkinesia and its molecular basis is not understood (Harper, 1991). In addition, the R6 mice appear to urinate more frequently as judged by wetting of the bedding. Urinary incontinence has also been noted in HD with symptoms including frequency, urgency, nocturia, and incontinence (Wheeler et al., 1985). Finally, chorea affecting face, jaw, and pharyngeal muscles affects both speech and swallowing and can also cause grunting and clicking sounds that may reflect respiratory movements (Harper, 1991). It is possible that the unusual vocalizations made by the R6 transgenes arise by a similar mechanism.

A landmark study of the neuropathology of HD has classified the neuropathological changes into five grades that progress from grade 0, in which HD brains show no gross or microscopic abnormalities consistent with HD despite premortem symptomatology and positive family history, to grade 4, in which the most extreme atrophy is observed (Vonsattel et al., 1985). The brains from the R6/2 transgenic mice were found to be on average 19% smaller than those of their normal littermates, a reduction in size that was maintained through all CNS structures. This finding is consistent with neuropathological changes occurring in HD in which it has been noted that a 30% reduction in brain weight in HD is associated with 20%–30% areal reductions in cerebral cortex, white matter, hippocampus, amygdala, and thalamus (de la Monte et al., 1988). This atrophy was similar for all grades of HD, suggesting that the shrinkage of these structures occurs early in the

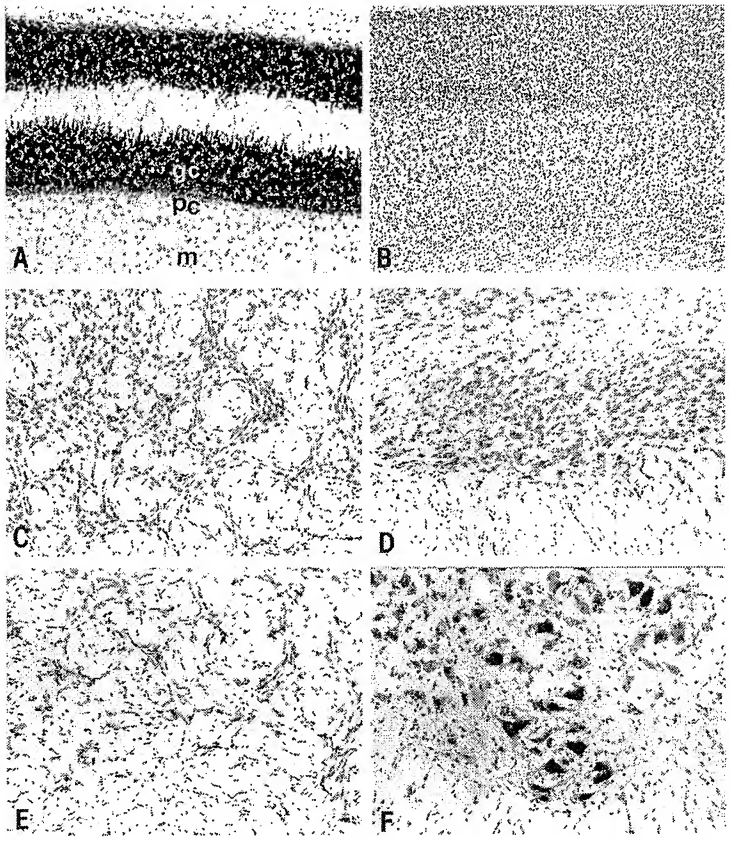


Figure 7. Nissl Sections throughout the CNS
Coronal sections through the cerebellum (A), cerebral cortex (B), globus pallidus (C), subthalamic nucleus (D), entopeduncular nucleus (E) and ventral horn of the lumbar spinal cord (F) of an R6/2 transgenic mouse. Within the cerebellum, note the normal density of the granule cell layer (gc), the monocellular layer of pyramidal cells (pc) and the normal structure of the molecular layer (m).

disease process, is not progressive, and reflects cell loss of both neurons as well as fibers. Interestingly, gliosis was not readily apparent in these structures, and the neuronal density was assessed to be normal (de la Monte et al., 1988). In contrast, a 60% reduction in the cross-sectional area of the caudate, putamen, and globus pallidus increases with the higher grades of HD brains, indicating that these structures progressively degenerate with prolongued survival. It is this specific progressive atrophy, associated with reactive astrocytosis, that was not apparent in the R6/2 transgenes and is also absent from grade 0 HD brains (Myers et al., 1991; Vonsattel et al., 1985). The grade 0 brains came from patients that had had HD symptomatology for between 2 and 13 years (Vonsattel et al., 1985; Myers et al., 1988; Hedreen and Folstein et al., 1995), thereby providing no pathological correlate for chorea and other early signs (Hedreen and Folstein et al., 1995). It seems likely that the brains of the R6/2 transgenes have neuropathology consistent with that found in the early stages of HD and that the progression of the phenotype in these mice is so rapid that there is insufficient time for the progressive atrophy to take place. A detailed morphometric analysis did uncover a neuronal loss in the caudate of grade 0 brains (Myers et al., 1991), and the absence of reactive astrocytosis was taken as evidence that the neuronal cell loss was not a recent event and may support the hypothesis that the HD striatum is compromised from early in development (Myers et al., 1991). A detailed morphometric analysis of the R6/2 transgene brains is merited. The neuropathological analysis of the transgenes was also focused on the additional regions that undergo neurodegeneration in the polygln expansion diseases as a whole, and no evidence of localized neurodegeneration was identified.

To date, five neurodegenerative diseases have been described that are caused by polygln expansions in ubiquitously expressed unrelated proteins. It is most probable that in each case the polygln expansion confers a gain of function to the proteins and that this may operate by a common molecular mechanism. It has been proposed that the specific selective cell death is directed by the remainder of the respective proteins. The R6 transgene protein products contain polygln tracts in a domain consisting of only 69 other amino acids amounting to ~3% of the htt protein. Therefore, the R6 transgenic mice might be expected to represent a generic CAG/gln disease model rather than a specific model of HD. However, the R6 mice do not develop a pronounced ataxia as described by Burright et al. (1995) and Ikeda et al. (1996). They do not develop a wide-based gait or fall while moving, are able to rear, and do not lose their righting response when turned onto their backs. This would suggest that there is no major cerebellar lesion and that the R6 lines do not display the major movement disorder of SCA1, SCA3, and late onset DRPLA. Similarly, they do not show a pronounced motorneuron disease, although the SBMA symptoms in

humans are mild with a very slow progression and it would probably be difficult to identify this component as part of the complex R6 phenotype. The diagnosis of HD and DRPLA was not infrequently confused before the advent of mutation analysis afforded an unequivocal test. Both disorders present with complex and variable symptoms that can include chorea, myoclonus, dystonia, dysarthria, and seizures. Some features are more or less associated with the juvenile or adult forms but the boundaries are not absolute. It would therefore be difficult to express any strong claims as to the specificity of a mouse model with respect to these two diseases.

The R6 mice are the first transgenic model of a polygln expansion disease in which the transgenes are ubiquitously expressed (as are the mutant human genes). The two previous reports of a neurological phenotype observed in mice transgenic for a protein carrying a polygln repeat expansion used a Purkinje cell specific promoter to drive either a *SCA1* cDNA construct with $(CAG)_{82}$ (Burright et al., 1995) or a $(CAG)_{79}$ polygln tract in isolation (Ikeda et al., 1996). Purkinje cell death was identified with a corresponding ataxic phenotype. It is possible that comparable overexpression of these constructs in any other cell would also demonstrate toxicity. The dramatic dosage effect on the phenotype observed with the R6 transgenes expressing at less than endogenous levels suggest that ubiquitous overexpression of the R6 transgene could be lethal.

The apparent absence of specific neurodegeneration in the R6 mice supports the possibility that localized atrophy may be secondary to a primary imbalance that is directly responsible for the clinical symptoms that arise in HD. Indeed, replication of the patterns of cell death observed in HD by intrastriatal injections of quinolinic acid does not cause chorea in rats (Harper, 1991). It remains remarkable that the introduction of the expanded version of the polygln-containing domain of htt protein into transgenic mice has succeeded in reproducing not only features of the movement disorder, but also other aspects of the complex HD phenotype.

Two further lines of transgenic mice are required to determine the extent to which the R6 mice represent a model of HD. First, mice transgenic for the entire *HD* gene carrying repeat expansions of a comparable size must be generated. The large size of the *HD* gene necessitates that the construct be introduced in the form of a YAC clone (experiments in progress). An identical phenotype would indicate that the remainder of the htt protein is superfluous to the course of the disease, and any differences would aid in the dissection of the protein into functional domains. Second, mice transgenic for the nonexpanded CAG repeat version of the R6 lines have not been described in this paper. The original purpose of the R6 transgenes was to study repeat stability and, consequently, the nonexpanded controls were not generated in parallel. However, it is important to characterize such mice, to rule out the unlikely scenario that the phenotype observed is the result of a novel peptide. Three founders have now been established that contain the SacI–EcoRI fragment with a $(CAG)_{18}$ tract: Hdex/6, Hdex/27, and Hdex/28. F1 mice derived from the Hdex/6 founder are currently 20 weeks, and the mice show no signs of a neurological phenotype or weight loss. These mice are twice as old as the R6/2 mice at the onset of the phenotype. Quantitative RNA analysis shows the Hdex/6 transgene to be expressed at levels comparable to that in the R6/2 and R6/5 lines; however, it is not possible to use the 1C2 antibody to detect the Hdex/6 transgene protein as this is specific to polyglutamine expansions. The Hdex lines will be bred to homozygosity and the mice observed over the course of at least one year.

This work raises the intriguing possibility that exon 1 of the *HD* gene carrying highly expanded repeats is sufficient to generate a transgenic model of HD. The mutation is predicted to operate by conferring a gain of function to the mutated protein to which some cells are particularly sensitive. The cell-selective toxicity may be afforded by differing compartmentalization of the polygln-carrying proteins or by the specificity of their intermolecular interactions. In order that the small R6 transgene could initiate a chain of molecular events comparable to those involving the entire htt protein, it would be necessary to predict that the transgene occupies the same subcellular localization. It has not been possible to make this comparison as our attempts at immunohistochemistry with the 1C2 antibody have been consistently unsuccessful, and in addition, the subcellular localization of htt remains to some extent controversial. If the selectivity of the cell death arises through the interacting proteins, the polygln-containing domain of the htt protein must be sufficient to convey this specificity. There may be some evidence to suggest that this could be the case, arising from the isolation of HAP1 (huntingtin associated protein 1) (Li et al., 1995). HAP1 binds to htt containing a polygln of 21 residues, and the association is enhanced by increasing lengths of the gln repeat. There was no binding to atrophin-1 (the mutant protein in DRPLA) also containing 21 glutamines.

It is impossible to predict the accuracy with which transgenic mouse lines will model a corresponding human disease. The R6 transgenes display many characteristics of HD, and had this phenotype arisen in mice transgenic for the entire mutant protein, the model would have needed little justification. It is clearly possible that the polygln-containing domain may be the only part of the htt protein involved in the disease process. The R6 transgenic mice already provide a valuable resource for uncovering the molecular pathology of HD and may present a target for the testing of potential therapeutic interventions.

Experimental Procedures

Genotyping

DNA was prepared from tail biopsy and Southern blots and hybridizations were as described (Monaco et al., 1985). CAG repeats were sized by PCR using FAM-labeled primer 31329 (ATGAAGGCCTTC GAGTCCCTCAAGTCCTTC) and primer 33934 (GGCGGCTGAG GAAGCTGAGGA) in AM buffer (67 mM Tris–HCl [pH 8.8], 16.6 mM NH_4SO_4, 2.0 mM $MgCl_2$ 0.17 mg/ml BSA, 10 mM 2-mercaptoethanol), 10% DMSO, 200 μM dNTPs, 8 ng/μl primers with 0.5 U/μl Taq polymerase (Cetus). Cycling conditions were 90" @ 94°C, 25 × (30" @ 94°C, 30" @ 65°C, 90" @ 72°C), 10' @ 72°C. PCR products were sized using an ABI sequencer and the Genescan and Genotyper software packages. The size of the CAG repeat was 85 bp less than the size of the PCR product.

RNA Analysis

Northern blots were prepared by standard methods and hybridized as described (Monaco et al., 1985). RNA was reverse transcribed (14 U/μl MMTV RTase, BRL) in 50 mM KCl, 10 mM Tris–HCl (pH 9.0), 0.1% Triton X-100, 6.5 mM $MgCl_2$, 10 mM DTT, 1 mM dNTPs, 10 ng/μl random hexamers with 0.35 U/μl RNasin (Promega) at 10' @ 23°C and then 40' @ 37°C. Primers for specific transgene RNA detection were 33935 (CGGCTGAGGCAGCAGCGGCTGT) and 35093 (GCAGCAGCAGCAGCAACAGCCGCCACCGCC). PCR was in AM buffer, 10% DMSO, 200 μM dNTPs, 10 ng/μl primer with 0.5 U/μl Taq polymerase (Cetus). Cycling conditions were 90" @ 94°C, 34 × (30" @ 94°C, 30" @ 68°C, 90" @ 72°C), 10' @ 72°C.

Protein Analysis

Frozen tissue was homogenized in 50–100 μl 50 mM Tris (pH 8.0), 150 mM NaCl, 1% NP-40, 0.5% Deoxycholate, 0.1% SDS, and 1 mM 2-mercaptoethanol with 1 mM PMSF, 0.5 mM DTT, 25 mM benzamidine and leupeptin, pepstatin and chymostatin each at 200 ng/ml. Homogenates were sonicated on ice 10–20 s, spun at high speed at 4°C, and the supernatant transferred to a fresh tube. Protein was quantified by the Bradford assay when in sufficient quantity. Approximately 50 μg of protein was loaded per track onto 6% or 10% SDS–PAGE gels. Kaleidoscope prestained standards were used as size markers (Biorad). Fibrinogen (Sigma) was added as a size marker of 330 kDa (Jou and Myers, 1995). Proteins were transferred to PVDF membranes (Biorad) that were blocked at 4°C overnight in PBS with 5% nonfat dry milk and 2% fetal calf serum. Immunoprobing with antibody 1C2 was at a 1:2000 dilution in PBS with 0.5% nonfat dry milk for 1 hr at RT. Washes were in PBS containing 1% NP-40 and 1% fetal calf serum. Secondary antibody probing and detection was by use of the ECL kit (Amersham).

Histopathology

Brains from nine R6/2 transgenes and nine nontransgenic littermates were analyzed for neuropathological change. A 1:3 series of sections was stained for Nissl substance with thionin, or processed free floating for the immunocytochemical localization of the glial marker, glial fibrillary acidic protein (GFAP), or the macrophage/microglial marker F4/80. Nuclear cells groups within the mouse brain were verified by reference to Sidman, Angevine, and Taber-Pierce (Sidman et al., 1971).

Acknowledgments

This paper is dedicated to the memory of Dennis Shea. The authors wish to thank Nancy Wexler and Anne Young for helpful discussions regarding the symptoms and progression of HD. We also thank Carl Hobbs for preliminary histopathology and Yuh-Shan Jou and Rick Myers for making their αHD1 antibody available. This work was supported by grants from the Medical Research Council, the Hereditary Disease Foundation (in the form of an award donated by Harry Liebermann), and the Special Trustees of Guy's Hospital.

Received July 25, 1996; revised September 10, 1996.

References

Burright, E.N., Clark, H.B., Servadio, A., Matilla, T., Fedderson, R.M., Yunis, W.S., Duvick, L.A., Zoghbi, H.Y., and Orr, H.T. (1995). SCA1 transgenic mice: a model for neurodegeration caused by an expanded CAG trinucleotide repeat. Cell *82*, 937–948.

de la Monte, S.M., Vonsattel, J.-P., and Richardson, E.P. (1988). Morphometric demonstration of atrophic changes in the cerebral cortex, white matter and neostriatum in Huntington's disease. J. Neuropath. Exp. Neurol. *47*, 516–525.

DiFiglia, M., Sapp, E., Chase, K., Schwarz, C., Meloni, A., Young, C., Martin, E., Vonstattel, J.-P., Carraway, R., Reeves, S.A., Boyce, F.M., and Aronin, N. (1995). Huntingtin is a cytoplasmic protein associated with vesicles in human and rat brain neurons. Neuron *14*, 1075–1081.

Durr, A., Stevanin, G., Cancel, G., Duyckaerts, C., Abbas, N., Didierjean, O., Chneiweiss, H., Benomar, A., Lyon-Caen, O., Julien, J., et al. (1996). Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular and neuropathological features. Ann. Neurol. *39*, 490–499.

Harper, P.S. (1991). Huntington's Disease (London: W.B. Saunders).

Hedreen, J.C., and Folstein, S.E. (1995). Early loss of early neostriatal neurons in Huntington's disease. J. Neuropath. Exp. Neurol. *54*, 105–120.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is unstable on Huntington's disease chromosomes. Cell *72*, 971–983.

Ikeda, H., Yamaguchi, M., Sugai, S., Aze, Y., Narumiya, S., and Kakizuka, A. (1996). Expanded polyglutamine in the Machado-Joseph disease protein induces cell death in vitro and in vivo. Nature Genet. *13*, 196–202.

Jou, Y.-S., and Myers, R.M. (1995). Evidence from antibody studies that the CAG repeat in the Huntington disease gene is expressed in the protein. Hum. Mol. Genet. *4*, 465–469.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S., and Kakizuka, A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nature Genet. *8*, 221–228.

Kennedy, W.R., Alter, M., and Sung, J.H. (1968). Progressive proximal spinal and bulbar atrophy of late onset. A sex linked recessive trait. Neurology *18*, 671–680.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., et al. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nature Genet. *6*, 9–13.

La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck, K.H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature *352*, 77–79.

Lauder, I.M., Martin, W.B., Gordon, E.D., Lawson, D.D., Campbell, R.S.F., and Watrach, A.M. (1954). A survey of canine distemper. Veterinary Record *66*, 607–611.

Li, X.-J., Li, S.-H., Sharp, A.H., Nucifora, F.C., Schilling, G., Lanahan, A., Worley, P., Snyder, S.H., and Ross, C.A. (1995). A huntingtin-associated protein enriched in brain with implications for pathology. Nature *378*, 398–402.

Lyon, M.F., and Searle, A.G. (1990). Genetic Variants and Strains of the Laboratory Mouse (Oxford: Oxford University Press).

Monaco, A.P., Bertelson, C.J., Middlesworth, W., Colletti, C.-A., Aldridge, J., Fischbeck, K.H., Bartlett, R., Pericak-Vance, M.A., Roses, A.D., and Kunkel, L.M. (1985). Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA probe. Nature *316*, 842–845.

Myers, R.H., Vonsattel, J.P., Stevens, T.J., Cupples, L.A., Richardson, E.P., Martin, J.B., and Bird, E.D. (1988). Clinical and neuropathological assessment of severity in Huntington's disease. Neurology *38*, 341–347.

Myers, R.H., Vonsattel, J.P., Paskevich, P.A., Kiely, D.K., Stevens, T.J., Cupples, L.A., Richardson, E.P., and Bird, E.D. (1991). Decreased neuronal and increased oligodendroglial densities in Huntington's disease caudate nucleus. J. Neuropath. Exp. Neurol. *50*, 729–742.

Orr, H.T., Chung, M., Banfi, S., Kwiatkowski, T.J., Jr., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P.W., and Zoghbi, H.Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature Genet. *4*, 221–226.

Sanberg, P.R., Fibiger, H.C., and Mark, R.F. (1981). Body weight and dietary factors on Huntington's disease patients compared with matched controls. Med. J. Aust. *1*, 407–409.

Servadio, A., Koshy, B., Armstrong, D., Antalffy, B., Orr, H.T., and Zoghbi, H.Y. (1995). Expression analysis of the ataxin-1 protein in tissues from normal and spinocerebelar ataxia type 1 individuals. Nature Genet. *10*, 94–98.

Sidman, R.L., Angevine, J.B.J., and Taber-Pierce, E. (1971). Atlas of the Mouse Brain and Spinal Cord (Cambridge, Massachusetts: Harvard University Press).

Strong, T.V., Tagle, D.A., Valdes, J.M., Elmer, L.W., Boehm, K., Swaroop, M., Kaatz, K.W., Collins, F.S., and Albin, R.L. (1993). Widespread expression of the human and rat Huntington's disease gene in brain and nonneuronal tissues. Nature Genet. *5*, 259–263.

Takahashi, H., Ohama, E., Naito, H., Takeda, S., Nakashima, S., Makifuchi, T., and Ikuta, F. (1988). Hereditary dentatorubral-pallidoluysian atrophy: clinical and pathological variants in a family. Neurology *38*, 1065–1070.

Telenius, H., Kremer, H.P.H., Theilmann, J., Andrew, S.E., Almquist, E., Anvret, M., Greenberg, C., Greenberg, J., Lucotte, G., Squitieri, F., Starr, E., Goldberg, Y.P., and Hayden, M.R. (1993). Molecular analysis of juvenile Huntington disease: the major influence on (CAG)n repeat length is the sex of the affected parent. Hum. Mol. Genet. *2*, 1535–1540.

Trottier, Y., Devys, D., Imbert, G., Sandou, F., An, I., Lutz, Y., Weber, C., Agid, Y., Hirsch, E.C., and Mandel, J.-L. (1995a). Cellular localisation of the Huntington's disease protein and discrimination of the normal and mutated forms. Nature Genet. *10*, 104–110.

Trottier, Y., Lutz, Y., Stevanin, G., Imbert, G., Devys, D., Cancel, G., Sandou, F., Weber, C., David, G., Tora, L., et al. (1995b). Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. Nature *378*, 403–406.

Vonsattel, J.-P., Myers, R.H., Stevens, T.J., Ferrante, R.J., Bird, E.D., and Richardson, E.P. (1985). Neuropathological classification of Huntington's disease. J. Neuropath. Exp. Neurol. *44*, 559–577.

Wheeler, J.S., Sax, D.S., Krane, R.J., and Siroky, M.B. (1985). Vesicourethral function in Huntington's chorea. Brit. J. Urol. *57*, 63–66.

Yazawa, I., Nukina, N., Hashida, H., Goto, J., Yamada, M., and Kanazawa, I. (1995). Abnormal gene product identified in hereditary dentatorubral-pallidoluysian atrophy (DRPLA) brain. Nature Genet. *10*, 99–103.

Zoghbi, H.Y. (1993). In Current Neurology, S.H. Appel, ed. (St. Louis, MO: Mosby–Year Book), pp. 87–110.